

Intrinsic Persistent Homology via Density-based Metric Learning

Ximena Fernández

Department of Mathematics

Swansea University, UK

Departamento de Matemática and IMAS-CONICET

FCEN, Universidad de Buenos Aires, Argentina

XFERNAND@DM.UBA.AR

Eugenio Borghini

Departamento de Matemática and IMAS-CONICET

FCEN, Universidad de Buenos Aires, Argentina

EBORGHINI@DM.UBA.AR

Gabriel Mindlin

IFIBA, CONICET and Departamento de Física

FCEN, Universidad de Buenos Aires, Argentina

GABO@DF.UBA.AR

Pablo Groisman

Departamento de Matemática and IMAS-CONICET

FCEN, Universidad de Buenos Aires, Argentina

PGROISMA@DM.UBA.AR

NYU-ECNU Institute of Mathematical Sciences

NYU Shanghai

Editor: Aryeh Kontorovich

Abstract

We address the problem of estimating topological features from data in high dimensional Euclidean spaces under the manifold assumption. Our approach is based on the computation of persistent homology of the space of data points endowed with a sample metric known as Fermat distance. We prove that such metric space converges almost surely to the manifold itself endowed with an intrinsic metric that accounts for both the geometry of the manifold and the density that produces the sample. This fact implies the convergence of the associated persistence diagrams. The use of this intrinsic distance when computing persistent homology presents advantageous properties such as robustness to the presence of outliers in the input data and less sensitiveness to the particular embedding of the underlying manifold in the ambient space. We use these ideas to propose and implement a method for pattern recognition and anomaly detection in time series, which is evaluated in applications to real data.

Keywords: topological data analysis, persistent homology, manifold learning, distance learning, time series

1. Introduction

1.1 Motivation and Problem Statement.

It is a common situation in machine learning that the given data represents a possibly noisy finite sample of a geometric object embedded in a high dimensional Euclidean space.

This is the case, for instance, in the analysis of time series arising from observations of a dynamical system, where a spatial representation of the data can be interpreted as a sample of a geometric structure — the *attractor* — encoding valuable information of the underlying system’s behaviour. Under the manifold assumption, both the metric and the density of the sample play a central role in the process of reconstruction of topological properties of the underlying shape.

From a theoretical point of view, the problem can be stated as follows. Let \mathbb{X}_n be a set of n sample points with common density f supported on a smooth compact Riemannian manifold \mathcal{M} embedded in \mathbb{R}^D . We are interested in recovering topological features of \mathcal{M} from the sample $\mathbb{X}_n \subseteq \mathbb{R}^D$ in a setting in which both \mathcal{M} and f are assumed to be unknown. A standard approach to accomplish this task consists in applying to \mathbb{X}_n a computational technique known as *persistent homology*, which allows to obtain qualitative information about connected components, cycles, voids and higher dimensional holes from the point cloud. Here, the sample \mathbb{X}_n is considered as a metric space endowed with some computable distance, such as the Euclidean distance or an estimator of the inherited geodesic distance. Although the topological information carried by \mathcal{M} remains the same when endowed with any Riemannian metric, the output of the application of persistent homology to \mathbb{X}_n strongly depends on the particular distance function employed. In this article, we consider a computable estimator defined over \mathbb{X}_n of a certain Riemannian metric on \mathcal{M} that takes into account the density f , which was called *Fermat distance* by Groisman et al. (2022). We show that the use of this density-based intrinsic metric in the computation of persistent homology can lead to results that overcome simultaneously certain weaknesses of standard approaches, such as the sensitivity to outliers and the dependence on the embedding of the sample in the ambient space.

Persistent homology is a central technique in Topological Data Analysis (TDA) developed to infer the *homology groups* of a space by studying a sample \mathbb{X}_n at *all* scales of resolution at the same time (see Edelsbrunner and Harer, 2008; Edelsbrunner et al., 2002; Boissonnat et al., 2018; Niyogi et al., 2008; Zomorodian and Carlsson, 2005). It has found applications in many fields, including neuroscience (Giusti et al., 2015), finance (Gidea and Katz, 2018), signal processing (Perea and Harer, 2015; Tralie and Perea, 2018; Perea, 2019), computational neural networks (Gabrielsson et al., 2020), virus evolution (Chan et al., 2013) and sensor networks (de Silva and Ghrist, 2007a). This method yields as output an object called *persistence diagram* associated to the sample. Under mild conditions, the homology groups of the underlying topological space can be read off the persistence diagram (see Edelsbrunner et al., 2002). Chazal et al. (2009, 2016) provided a general framework that allows to define persistence diagrams for infinite metric spaces instead of just finite approximations (samples). Thus, one can view the persistence diagram associated to a sample of a space as an estimate of a limiting object, namely, the persistence diagram of the entire space. When the distinction is needed, we will call these diagrams *sample persistence diagram* and *population persistence diagram* respectively.

Our main result states that, under reasonable conditions, there is convergence as metric spaces of the sample \mathbb{X}_n endowed with a computable estimator of the Fermat distance towards the manifold \mathcal{M} (equipped with the Fermat distance) in the sense of Gromov–Hausdorff as the size n grows. When combined with the well-known *stability theorem* (Cohen-Steiner et al., 2007; Chazal et al., 2014, 2009), this approximation result as metric

spaces allows to deduce the convergence of the corresponding persistence diagrams. For this purpose, the space of diagrams is naturally equipped with the *bottleneck distance*. Approximation results that include convergence rates and confidence regions have been established when the metric of the target space is known; see e.g. (Fasy et al., 2014) where the Euclidean distance is considered for both the samples and the space and also (Chazal et al., 2015), where a general metric is used but assumed to be known in advance.

Persistence diagrams are known to be sensitive to the presence of outliers (see Bendich et al., 2011; Chazal et al., 2011; Buchet et al., 2016; Anai et al., 2019). Chazal et al. (2011); Anai et al. (2019) proposed filtrations of point clouds regarded as empirical measures in the ambient Euclidean space — called DTM-filtrations — to achieve a robust computation of ambient persistent homology. This theory was later extended to general metric spaces by Buchet et al. (2016). On the other hand, intrinsic versions of the classical Čech and Vietoris–Rips filtrations were developed with the aim of capturing topological properties of manifolds sitting in an Euclidean space which are independent of the embedding. The approach exhibited in this article handles both difficulties at the same time. Indeed, we show that sample persistence diagrams computed using the estimator of the (intrinsic) Fermat distance are both robust to outliers for positive degree and display the correct homology of the manifold for a longer parameter interval as compared with the use of ambient Euclidean distance.

We refer the reader to the video by Fernandez (2021b) for an introductory exposition of the contents of this article.

1.2 Contributions

Let (\mathcal{M}, ρ) be a smooth d -dimensional Riemannian manifold embedded in \mathbb{R}^D with density $f : \mathcal{M} \rightarrow \mathbb{R}_{>0}$ and a Riemannian density-based distance ρ (mainly, it will be the Fermat distance $d_{f,p}$ defined below).

For $p > 1$, the *population Fermat distance* is defined as

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

Here $x, y \in \mathcal{M}$, $|\cdot|$ denotes the Euclidean distance and the infimum is taken over all piecewise smooth curves $\gamma : I = [0, 1] \rightarrow \mathcal{M}$ with $\gamma(0) = x$, and $\gamma(1) = y$. In the special case when f is uniform, the population Fermat distance reduces to (a multiple of) the inherited Riemannian distance $d_{\mathcal{M}}$ from the ambient Euclidean space. When this is not the case, this distance takes into account the density, which may be advantageous in certain situations, like in the case of estimation of the topology of \mathcal{M} from samples with presence of noise and outliers. This metric was also considered in the works by Hwang et al. (2016); Mckenzie and Damelin (2019); Sapienza et al. (2018); Groisman et al. (2022).

Given a finite set of points \mathbb{X}_n , the *sample Fermat distance* between x, y 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 $\gamma = (x_0, x_1, \dots, x_{r+1})$ with $x_0 = x$, $x_{r+1} = y$ and $\{x_1, x_2, \dots, x_r\} \subseteq \mathbb{X}_n$.

Our main result states the Gromov–Hausdorff convergence (a.s.) of the sample endowed with the sample Fermat distance, appropriately re-scaled, to $(\mathcal{M}, d_{f,p})$.

Theorem *Let \mathcal{M} be a smooth, closed d -dimensional Riemannian manifold embedded in \mathbb{R}^D . Let $f : \mathcal{M} \rightarrow \mathbb{R}_{>0}$ be a smooth density function. Let $\mathbb{X}_n = \{x_1, x_2, \dots, x_n\} \subseteq \mathcal{M}$ be a set of n independent sample points in \mathcal{M} with common density f . Given $p > 1$, there exists a constant $\mu = \mu(p, d)$ such that for every $\lambda \in ((p-1)/pd, 1/d)$ and $\varepsilon > 0$ there exist $\theta > 0$ satisfying*

$$\mathbb{P}\left(d_{GH}\left((\mathcal{M}, d_{f,p}), (\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n, p})\right) > \varepsilon\right) \leq \exp\left(-\theta n^{(1-\lambda d)/(d+2p)}\right)$$

for n large enough, where d_{GH} stands for the Gromov-Hausdorff distance between metric spaces.

As a consequence of this result and the stability theorem for persistence diagrams we deduce the following convergence result.

Corollary *Let $\varepsilon > 0$ and $\lambda \in ((p-1)/pd, 1/d)$. There exists a constant $\theta > 0$ such that*

$$\mathbb{P}\left(d_b(\text{dgm}(\text{Filt}(\mathcal{M}, d_{f,p})), \text{dgm}(\text{Filt}(\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n, p}))) > \varepsilon\right) \leq \exp\left(-\theta n^{(1-\lambda d)/(d+2p)}\right)$$

for n large enough.

Here $\text{Filt}(\cdot)$ denotes either the Vietoris–Rips or Čech filtration, $\text{dgm}(\cdot)$ the associated persistence diagram and d_b is the bottleneck distance (see Section 3 for precise definitions). Since $(\mathcal{M}, d_{f,p})$ is a Riemannian manifold, its population persistence diagram $\text{dgm}(\text{Filt}(\mathcal{M}, d_{f,p}))$ displays the correct homology up to the convexity radius $\text{conv}(\mathcal{M}, d_{f,p})$. In contrast, for $(\mathcal{M}, |\cdot|)$ this is guaranteed only up to the reach $\tau_{\mathcal{M}}$. It is easy to find examples of manifolds in which $\text{conv}(\mathcal{M}, d_{f,p})$ is much larger than $\tau_{\mathcal{M}}$.

On the other hand, we prove that for a reasonable upper bound r on the filtration parameter, $\text{dgm}(\text{Rips}_{<r}(\mathbb{X}_n, d_{\mathbb{X}_n, p}))$ is robust to outliers for homology degree greater than 0.

Proposition *Let \mathbb{X}_n be a sample of \mathcal{M} and let $Y \subseteq \mathbb{R}^D \setminus \mathcal{M}$ be a finite set of outliers. Let $\delta = \min \left\{ \min_{y \in Y} d_E(y, Y \setminus \{y\}), d_E(\mathbb{X}_n, Y) \right\}$, where d_E denotes the Euclidean distance between sets. For all $k > 0$ and $p > 1$,*

$$\text{dgm}_k(\text{Rips}_{<\delta^p}(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})) = \text{dgm}_k(\text{Rips}_{<\delta^p}(\mathbb{X}_n, d_{\mathbb{X}_n, p})),$$

where $\text{Rips}_{<\delta^p}$ stands for the Rips filtration up to parameter δ^p and dgm_k for the persistent homology of degree k .

The threshold δ^p is restrictive if it is below $\text{diam}(\mathbb{X}_n, d_{\mathbb{X}_n, p})$. However, we will show that under a natural model for the outliers, $\delta^p > \text{diam}(\mathbb{X}_n, d_{\mathbb{X}_n, p})$ for large enough p .

1.3 Applications to Signal Analysis.

The study of time series — specially, derived from dynamical systems — through the inference of homology groups of a certain associated space called *delay embedding* was pioneered in the works by Perea (2019); Perea and Harer (2015). The construction of the delay embedding of a time series heavily depends on the dimension or number of independent variables

of the underlying system, and the choice of a parameter called *time delay*. It often leads to analyse subspaces of a sufficiently high dimensional Euclidean space, which makes the inference of topological information unstable.

In first place, by means of concrete examples involving the Lorenz attractor and noisy periodic signals, we show that the use of Fermat distance in this method can lead to a more robust inference of the delay embeddings' topological features. The reason behind this is that the Fermat distance is less prone, compared to the Euclidean distance, to the effect known as *curse of dimensionality* and less dependent on the particular embedding. We also describe a method to detect change-points in the time series through the study of the evolution in time of the persistence diagrams of the corresponding time-delay embeddings. This is applied to discover anomalies in electrocardiogram signals and different patterns in the song of canaries corresponding to different syllables.

The code to replicate the computational examples and applications can be found at the repository by Fernandez (2021a).

1.4 Related Work

The sample Fermat distance was introduced independently in the articles by Sapienza et al. (2018); Mckenzie and Damelin (2019). The study of approximations of density based metric from samples was suggested by Vincent and Bengio (2003) and developed by Sajama and Orlitsky (2005). Cohen et al. (2015); Chu et al. (2020) analyzed a general family of metrics that includes the population Fermat distance and deeply studied the case $p = 2$ of sample Fermat distance, which was also called power weighted shortest distance by Mckenzie and Damelin (2019). Groisman et al. (2022) proved that it is possible to recover the population Fermat distance $d_{f,p}$ for d -dimensional manifolds which are isometrically embedded (closures of) open sets of \mathbb{R}^d in \mathbb{R}^D as the limit of the sample Fermat distance. In the related work by Hwang et al. (2016) it was shown that in the same context, a statistic that is similar to the sample Fermat distance but uses the inherited Riemannian distance d_M between consecutive points in a path instead of the Euclidean one to measure its cost, also converges almost surely to the Fermat distance. We remark that this statistic cannot be computed from the sample since the inherited distance is not assumed to be known in advance. However, the results by Hwang et al. (2016) provides an essential and strong foundation on the basis of which our main result is built over.

The problem of learning geodesic distances from samples for submanifolds of the Euclidean space, specially with the aim of reducing dimensionality and visualizing data, has a long history; see for instance (Tenenbaum et al., 2000; Mémoli and Sapiro, 2005). On the other hand, the problem of estimating the persistence diagram of a submanifold of an Euclidean space from a sample has been studied by Fasy et al. (2014); Chazal et al. (2015), where the underlying metric is assumed to be known. In this setting, both Chazal et al. (2015) and Fasy et al. (2014) were able to prove the following satisfying result: the persistence diagrams computed using the sample converge almost surely (in the sense of bottleneck distance) to the persistence diagram of the desired metric space. Moreover, they gave exponentially small bounds in the size of the sample for the probability of the bottleneck distance between the corresponding persistence diagrams being larger than some positive number; see (Chazal et al., 2015, Corollary 3) and (Fasy et al., 2014, Lemma 4),

where in addition confidence sets for persistence diagrams are provided. In a different direction, the advantages of computing persistence diagrams of submanifolds of an Euclidean space using alternative metrics — more specifically, metrics based on diffusion geometry and random walks — were explored experimentally by Bendich et al. (2011).

1.5 Structure of the Paper

In Section 2 we prove our main result Theorem 7 regarding the Gromov–Hausdorff convergence of metric spaces using, respectively, the sample and the population Fermat distance. Section 3 includes an introduction to persistent homology and is devoted to the study of persistence diagrams of manifolds endowed with Fermat distance. We deduce in first place the convergence of sample persistence diagrams to population persistence diagrams. Then, we show that by using these intrinsic metrics the topological features last longer in the persistence diagrams. Finally, we show that Fermat-based persistence diagrams are robust to the presence of outliers for positive homology degree. In Section 4 we present a method for pattern recognition in time series, which is applied to real data from electrocardiograms and songs of canaries. Appendix A contains the proofs of some technical results (Proposition 5 and Lemma 8), required as intermediate steps to prove Theorem 7.

2. Density-based Distance Learning

In this section we prove the main theorem of the article, which states that the sample \mathbb{X}_n , considered as a metric space with the sample Fermat distance (appropriately re-scaled), converges almost surely to $(\mathcal{M}, d_{f,p})$ in the sense of Gromov–Hausdorff.

We begin by introducing the *population Fermat distance* for a smooth closed Riemannian manifold without boundary \mathcal{M} of dimension $d > 1$ with Riemannian metric tensor g together with a positive C^∞ density function $f : \mathcal{M} \rightarrow \mathbb{R}_{>0}$. For $p > 1$, consider the deformed metric tensor $g_p = f^{2(1-p)/d} g$ given by a conformal transformation of the original metric g . Since f is smooth, g_p is a Riemannian metric tensor. Thus, \mathcal{M} has a metric space structure given by the geodesic distance with respect to g_p , denoted by $d_{f,p}$.

Definition 1 (Hwang et al., 2016) *For $p > 1$, the population Fermat distance between $x, y \in \mathcal{M}$ is defined as*

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

where the infimum is taken over all piecewise smooth curves $\gamma : I \rightarrow \mathcal{M}$ with $\gamma_0 = x$, and $\gamma_1 = y$.

Notice that geodesics in \mathcal{M} with respect to the distance $d_{f,p}$ are more likely to lie in regions with high values of f . The name *Fermat distance* comes from the analogy with optics, in which $d_{f,p}$ is the optical distance as defined by Fermat's principle when the refraction index is given by $f^{-(p-1)/d}$.

Consider now a set $\mathbb{X}_n = \{x_1, x_2, \dots, x_n\} \subseteq \mathcal{M}$ of n sample points in \mathcal{M} with common density f . Suppose that \mathcal{M} is embedded in \mathbb{R}^D and it is endowed with the standard inherited Riemannian metric. Our aim is to approximate $d_{f,p}(x, y)$, assuming no knowledge about \mathcal{M}

and the Riemannian distance defined on it. To achieve this, we will define an estimator for this distance over the sample. We denote by $|x - y|$ the Euclidean distance between points $x, y \in \mathcal{M}$.

Definition 2 (*Sapienza et al., 2018; Mckenzie and Damelin, 2019*) For $p > 1$, the sample Fermat distance between $x, y \in \mathcal{M}$ 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 $\gamma = (x_0, x_1, \dots, x_{r+1})$ of finite length with $x_0 = x$, $x_{r+1} = y$ and $\{x_1, x_2, \dots, x_r\} \subseteq \mathbb{X}_n$.

Since $p > 1$, geodesics with respect to this distance are also likely to lie in regions with high density of points in \mathbb{X}_n . This is due to the fact that paths with short edges are favored even if they have large total (Euclidean) length.

We remark here that, for technical reasons, we adopt a slightly different definition for the sample Fermat distance than the original one from Sapienza et al. (2018). Namely, in the original setting, only paths completely contained in \mathbb{X}_n are considered, including the endpoints. Points that are not in the sample \mathbb{X}_n are projected to the nearest point in \mathbb{X}_n . In consequence, our sample Fermat distance here does not generally induce a pseudometric over \mathcal{M} , but only a metric when restricted to \mathbb{X}_n .

Example 1 (Eyeglasses) The effect of taking different values of p for the sample Fermat distance $d_{\mathbb{X}_n, p}$ in the geometry of a manifold is illustrated below. Concretely, the eyeglasses curve in \mathbb{R}^2 uniformly sampled and perturbed with Gaussian noise is considered (see Figure 1). We compute the sample Fermat distance between each pair of points for a series of values of $p > 1$ and embed the sampled points in \mathbb{R}^2 in such a way that the Euclidean distance in the embedding reflects the Fermat distance, using the Multidimensional Scaling algorithm (MDS). As p becomes larger, the geometry of the data overcomes the bottleneck region and it deforms into a circle. We also compute the Isomap embedding in \mathbb{R}^2 posed by Bernstein et al. (2000). Recall that the Isomap embedding is the MDS projection with an estimator of the inherited Riemannian distance based in the k -NN graph as input distances (see Bernstein et al., 2000, Section 5). Due to the noise near the bottleneck region, some points that are far in the sense of the inherited Riemannian distance become close in the distance estimated from the k -NN graph. Note that Isomap embedding is highly sensitive to noise, while with Fermat distance the points lying in low density regions are mapped to points that are far from the rest of the sample. The larger the power p , the stronger this effect. This feature allows Fermat distance to reconstruct the underlying topology of the manifold in the present case, even with noise, for a range of values of p .

Remark 3 (The role of p) The parameter p in the definition of the population Fermat distance $d_{f,p}$ controls the density weight $f^{-(p-1)/d}$ in the computation of geodesics. Whereas for $p = 1$ the optimal paths are obtained in classic geodesic paths, for large p they might significantly differ, being mostly restricted to areas of high density. In practice, the value of p in the sample Fermat distance $d_{\mathbb{X}_n, p}$ quantifies the balance between the embedding and the

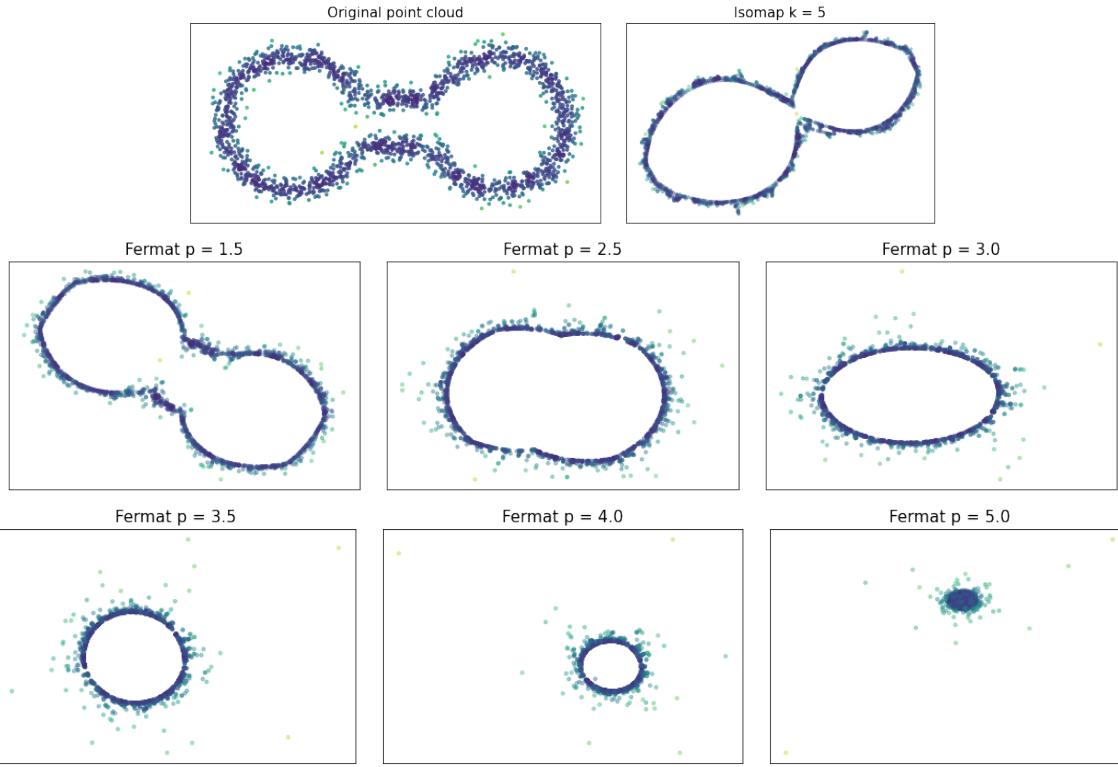


Figure 1: Top: A sample with noise of 2000 points of the eyeglasses dataset and Isomap projection with $k = 5$ (similar results are obtained for all reasonable values of k). Points are coloured according to local density. Middle and bottom: MDS embedding in \mathbb{R}^2 using Fermat distance for different values of p .

density of a given sample \mathbb{X}_n when estimating the optimal paths (notice that it is equivalent to the Euclidean distance for $p = 1$). In general, there is a reasonable large — although bounded — interval of values of p for which the estimator $d_{\mathbb{X}_n, p}$ allows to recover the intrinsic geometry of the sample \mathbb{X}_n even in presence of noise (c.f. Example 1). A similar phenomena can be experimentally observed when it is used in clustering tasks, as shown in simulations by Sapienza et al. (2018) and Little et al. (2022).

Remark 4 (Dimensionality reduction) The estimation of Fermat distance on input data, when coupled with the MDS projection, produces a new method to achieve dimensionality reduction. This strategy is in analogy with the popular algorithm Isomap by Tenenbaum et al. (2000). It is known that Isomap suffers from topological instability in presence of noise, since it may construct erroneous connections (called short-circuits) in the k -NN graph that potentially impair its performance (see Balasubramanian and Schwartz (2002)). In contrast, since noise generally corresponds with regions of low density, noisy points are treated by our method almost as not being part of the manifold. These effects increase with the value of p , and they might be advantageous for the inference of the right geometry of the data (c.f. Section 3.3).

Our first result, Proposition 5, shows that the sample Fermat distance converges to the population Fermat distance for closed (i.e. compact and without boundary) submanifolds of \mathbb{R}^D . A related result was previously proved by Groisman et al. (2022) for isometrically embedded (closures of) open sets of \mathbb{R}^d . Here we extend the class of manifolds to any compact manifold without boundary embedded in \mathbb{R}^D . Moreover, Proposition 5 states a *uniform* convergence for *any* two points in the manifold — not only pointwise, as stated by Groisman et al. (2022) —. This feature is essential to study both the manifold and the sample endowed with the (population and sample respectively) Fermat distance as *single* objects (metric spaces) and to prove convergence in the sense of Gromov–Hausdorff.

Let us fix some notations and general hypotheses. Hereafter, \mathcal{M} will denote a smooth d -dimensional closed Riemannian submanifold of \mathbb{R}^D endowed with the inherited Riemannian distance $d_{\mathcal{M}}$. We will consider a set $\mathbb{X}_n \subseteq \mathcal{M}$ of n independent random points with common smooth density $f: \mathcal{M} \rightarrow \mathbb{R}_{>0}$. We will denote by M_f and m_f the maximum and minimum values attained by f on \mathcal{M} , respectively. Observe that $0 < m_f < M_f < \infty$. Finally, given $p > 1$ we set $\alpha = 1/(d + 2p)$.

Proposition 5 *For every $p > 1$ and $\lambda \in ((p - 1)/pd, 1/d)$, given $\varepsilon > 0$ there exist $\mu, \theta > 0$ such that*

$$\mathbb{P} \left(\sup_{x,y} \left| n^{(p-1)/d} d_{\mathbb{X}_n,p}(x, y) - \mu d_{f,p}(x, y) \right| > \varepsilon \right) \leq \exp(-\theta n^{(1-\lambda d)\alpha})$$

for n large enough. The supremum is taken over $x, y \in \mathcal{M}$.

The constant μ from the statement is fixed throughout this manuscript and depends only on p and d . It was originally defined in (Howard and Newman, 1997, Lemma 3). The constant θ depends on ε, p, f and \mathcal{M} .

Proposition 5 is derived from a related result by Hwang et al. (2016), in which the authors establish the convergence of a sample statistic known as the *power-weighted shortest path* to the population Fermat distance. For $p > 1$ and points $x, y \in \mathcal{M}$, the power-weighted shortest path between x, y is defined as

$$L_{\mathbb{X}_n,p}(x, y) = \inf_{\gamma} \sum_{i=0}^k d_{\mathcal{M}}(x_{i+1}, x_i)^p \quad (1)$$

where the infimum is taken over all paths $\gamma = (x_0, \dots, x_{k+1})$ in \mathbb{X}_n of finite length with $x_0 = x, x_{k+1} = y$.

Theorem 6 (Hwang et al., 2016, Theorem 1) *Let $p > 1$ and $\varepsilon > 0$. Suppose that $(b_n)_{n \geq 1}$ is a sequence of positive real numbers such that $\frac{\log(n)}{nb_n^d} \rightarrow 0$ as n goes to infinity. Then, there exists a constant $\theta > 0$ (which depends on ε) such that*

$$\mathbb{P} \left(\sup_{\substack{x,y \in \mathcal{M} \\ d_{\mathcal{M}}(x,y) \geq b_n}} \left| \frac{n^{(p-1)/d} L_{\mathbb{X}_n,p}(x, y)}{d_{f,p}(x, y)} - \mu \right| > \varepsilon \right) \leq \exp(-\theta(nb_n^d)^\alpha)$$

for all sufficiently large n , where the supremum is taken over $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq b_n$.

As explained in the paragraph following Theorem 1 in (Hwang et al., 2016, p. 2793), the requirement that $\frac{\log(n)}{nb_n^d} \rightarrow 0$ is necessary in order to obtain a nontrivial upper bound for the probability.

Note that in Theorem 6, the convergence holds for the set of points $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y)$ greater than some sequence (b_n) . However, since we will be interested in studying the Gromov–Hausdorff convergence of the associated metric spaces (see (2) below), it is necessary to have uniform control of the convergence of the estimated distance for *all* points in the manifold. The uniform convergence is one of the main improvements upon Theorem 6 we show in Proposition 5. Also, notice that the proposed statistic $L_{\mathbb{X}_n, p}$ of $d_{f,p}$ is based on the previous knowledge of the inherited Riemannian distance $d_{\mathcal{M}}$. In the general data analysis setting, only a sample of points in a Euclidean space is given. Under the assumption that points lie on an (unknown) manifold \mathcal{M} , the goal is to find an estimator of the intrinsic distance $d_{f,p}$ that can be completely computed from the sample. In Proposition 5, we prove that sample Fermat distance $d_{\mathbb{X}_n, p}$ is indeed a good estimator of $d_{f,p}$.

Proposition 5 arises as a natural continuation of Theorem 6. The main idea of the proof is to show that any segment that is part of any shortest path with respect to $d_{\mathbb{X}_n, p}$ will be arbitrarily small with high probability if n is large enough. This will allow us to deduce that the power-weighted distance is well approximated by the sample Fermat distance. We defer the proof to Appendix A.

We will next estimate the Gromov–Hausdorff distance between the metric space \mathbb{X}_n with an appropriate re-scaling of the sample Fermat distance $d_{\mathbb{X}_n, p}$ and \mathcal{M} endowed with the population Fermat distance $d_{f,p}$. Recall that the *Gromov–Hausdorff distance* d_{GH} is a metric on the (isometry classes of) compact metric spaces that, roughly speaking, quantifies how difficult it is to match every point of a metric space $(\mathbb{X}, \rho_{\mathbb{X}})$ with some point of another space $(\mathbb{Y}, \rho_{\mathbb{Y}})$. More formally, it is defined as

$$d_{GH}((\mathbb{X}, \rho_{\mathbb{X}}), (\mathbb{Y}, \rho_{\mathbb{Y}})) := \inf\{d_H(h_1(\mathbb{X}), h_2(\mathbb{Y}))\}, \quad (2)$$

where the infimum is over all the isometric embeddings $h_1: \mathbb{X} \rightarrow \mathbb{W}$, $h_2: \mathbb{Y} \rightarrow \mathbb{W}$ in a common metric space \mathbb{W} and d_H stands for the Hausdorff distance. We will employ the following equivalent characterization of the Gromov–Hausdorff distance, which is often more convenient:

$$d_{GH}((\mathbb{X}, \rho_{\mathbb{X}}), (\mathbb{Y}, \rho_{\mathbb{Y}})) = \frac{1}{2} \inf_R \sup_{(x,y),(x',y') \in R} |\rho_{\mathbb{X}}(x, x') - \rho_{\mathbb{Y}}(y, y')|, \quad (3)$$

where the infimum is taken over subsets $R \subseteq \mathbb{X} \times \mathbb{Y}$ such that the projections $\pi_{\mathbb{X}}(R) = \mathbb{X}$, $\pi_{\mathbb{Y}}(R) = \mathbb{Y}$.

We are now ready to state our main theorem. For notational convenience, we set $d_{n,p} = \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n, p}$, the re-scaled sample Fermat distance on \mathbb{X}_n .

Theorem 7 *Let $\varepsilon > 0$ and $\lambda \in ((p-1)/pd, 1/d)$. There exists a constant $\theta > 0$ such that*

$$\mathbb{P}(d_{GH}((\mathcal{M}, d_{f,p}), (\mathbb{X}_n, d_{n,p})) > \varepsilon) \leq \exp(-\theta n^{(1-\lambda d)\alpha})$$

for n large enough and $\alpha = 1/(d + 2p)$.

Before presenting the proof of Theorem 7, we will need a preliminary lemma which asserts that, with high probability, no point of \mathcal{M} is too far from the nearest point of the sample. The argument of this proof is standard, but we include it in Appendix A for the reader's convenience.

Lemma 8 *For any $\kappa > 0$, the event*

$$\left\{ \sup_{x \in \mathcal{M}} d_{\mathcal{M}}(x, \mathbb{X}_n) \geq n^{(\kappa-1)/d} \right\}$$

holds with probability at most $\exp(-\theta n^\kappa)$ for some constant $\theta > 0$ if n is large enough.

We are now in position to prove Theorem 7.

Proof [Theorem 7] In order to compute the Gromov–Hausdorff distance between $(\mathcal{M}, d_{f,p})$ and $(\mathbb{X}_n, d_{n,p})$, we consider in (3) the relation

$$R = \{(x_i, x_i) : x_i \in \mathbb{X}_n\} \cup \{(x_y, y) : y \in \mathcal{M}, d_{f,p}(x_y, y) = d_{f,p}(\mathbb{X}_n, y)\}.$$

By a simple application of the triangle inequality we get that

$$d_{GH}((\mathcal{M}, d_{f,p}), (\mathbb{X}_n, d_{n,p})) \leq \frac{1}{2} \left(\sup_{x,y \in \mathbb{X}_n} |d_{f,p}(x, y) - d_{n,p}(x, y)| + 2 \sup_{y \in \mathcal{M}} d_{f,p}(\mathbb{X}_n, y) \right). \quad (4)$$

Observe that the two terms on the right hand side of the previous inequality can be bounded above by Proposition 5 and Lemma 8 respectively.

Given $\varepsilon > 0$, by (4) we have that

$$\begin{aligned} & \mathbb{P}\left(d_{GH}((\mathcal{M}, d_{f,p}), (\mathbb{X}_n, d_{n,p})) > \varepsilon/2\right) \\ & \leq \mathbb{P}\left(\sup_{x,y \in \mathbb{X}_n} |d_{f,p}(x, y) - d_{n,p}(x, y)| > \varepsilon/2\right) + \mathbb{P}\left(\sup_{y \in \mathcal{M}} d_{f,p}(\mathbb{X}_n, y) > \varepsilon/4\right) \end{aligned}$$

To bound the first term, we apply Proposition 5 to get

$$\mathbb{P}\left(\sup_{x,y \in \mathbb{X}_n} |d_{f,p}(x, y) - d_{n,p}(x, y)| > \varepsilon/2\right) \leq \exp(-\theta n^{(1-\lambda d)\alpha}).$$

for some positive constant θ and n sufficiently large. As for the second term, notice that since

$$d_{f,p}(x, y) \leq m_f^{-(p-1)/d} d_{\mathcal{M}}(x, y),$$

Lemma 8 implies

$$\mathbb{P}\left(\sup_{y \in \mathcal{M}} d_{f,p}(\mathbb{X}_n, y) > n^{(\alpha-1)/d} m_f^{(p-1)/d}\right) \leq \exp(-\theta n^\alpha)$$

for n large. The proof follows by noticing that the sequence $n^{(\alpha-1)/d} m_f^{-(p-1)/d}$ converges to 0 as n goes to infinity. \blacksquare

Remark 9 (Rate of convergence) The rate of convergence in Theorem 7 is related to the fluctuations of $n^{\frac{p-1}{d}} d_{\mathbb{X}_n, p}(x, y)$ around $\mu d_{f,p}(x, y)$ or, more coarsely, the variance of $n^{\frac{p-1}{d}} d_{\mathbb{X}_n, p}(x, y)$ (Damron and Wang (2016) provide strong evidence that the bias can be bounded by the variance). It is expected that this variance decreases as a power of n , i.e.

$$cn^{-\zeta} \leq \text{Var}\left(n^{\frac{p-1}{d}} d_{\mathbb{X}_n, p}(x, y)\right) \leq Cn^{-\zeta}$$

for a dimension-dependent constant $\zeta = \zeta(d) > 0$. The precise value of $\zeta(d)$ is a still open problem in probability theory in the context of First Passage Percolation (Howard and Newman (2001); Auffinger et al. (2017)). For $d = 1$ it can be proved that $\zeta = 1$. For $d \geq 2$ it is widely believed (Auffinger et al., 2017) that the exponent should not depend on p and that for $d = 2$ we should have $\zeta(2) = 2/3$. For $d \geq 3$ it is not clear what the value of $\zeta(d)$ should be. If we write $\zeta(d) = -2(\chi(d) - 1)/d$, it is expected that $\chi(d)$ should decrease with the dimension but there is not agreement on whether there exists some critical dimension d_c such that $\chi(d) = 0$ for $d \geq d_c$ or even if we should have $\chi(d) \rightarrow 0$ as $d \rightarrow \infty$ (Auffinger et al., 2017, Section 3). In Howard and Newman (2001) non-optimal rigorous bounds have been proven for Euclidean First Passage Percolation that in our context read

$$\mathbb{P}\left(d_{GH}\left((\mathcal{M}, d_{f,p}), (\mathbb{X}_n, \frac{n^{(p-1)/d}}{\mu} d_{\mathbb{X}_n, p})\right) > n^{-\frac{1}{d} + \varepsilon}\right) \leq C_1 \exp(-C_0 n^\varepsilon)$$

for positive constants C_0, C_1 depending on $\varepsilon > 0$. This bound follows immediately in our case when \mathcal{M} is the closure of a bounded open and convex set and f is constant on \mathcal{M} . For the general case considered in this manuscript we expect to have similar bounds. Obtaining those bounds would be highly valuable, but its analysis is out of the scope of this paper. We refer the reader to Little et al. (2022) for a detailed discussion about the rate of convergence.

3. Fermat-based Persistent Homology

In this section we explore the use of Fermat distance as input in the computation of the persistence diagram associated to a sample of a manifold. We deduce the almost sure convergence of persistence diagrams of the sample \mathbb{X}_n with the (re-scaled) sample Fermat distance towards the persistence diagram of $(\mathcal{M}, d_{f,p})$. We also show that we expect to read the correct homology of \mathcal{M} for a longer parameter interval in the diagram associated to the sample \mathbb{X}_n computed with Fermat distance as compared with the use of Euclidean distance. Finally, we prove that Fermat-based persistence diagrams are robust to the presence of outliers for homology degree greater than 0.

3.1 Convergence of Persistence Diagrams

We start by briefly recalling the main concepts and results in persistent homology theory and refer the reader to the works by Chazal et al. (2014, 2016) for a more thorough exposition.

For the computation of the persistent homology of a point cloud, one imagines each point as a *ball* (that is, representing a small surrounding region) and builds a combinatorial model for the space connecting the points according to whether the corresponding regions intersect. More precisely, for every fixed value of a parameter or *scale* that controls the size

of the region that each point represents, one gets a *simplicial complex* (i.e., a higher dimensional analogue of a graph). This family of simplicial complexes, also known as a *filtration*, is the input of the procedure to compute persistent homology. Indeed, the topological features of this family of complexes change as the scale parameter grows: different connected components join in one, some loops are filled, new cavities appear, etc. By analyzing these transitions, we are able to assign a *birth* and a *death* value to each of these features, and the difference between them represents its *persistence*. The most persistent features represent *topological signatures*, whereas the shortest intervals may be considered as *noise*. The output of this procedure is summarized in an object called *persistence diagram*. We next give the formal definitions.

Given a (possibly infinite) metric space (\mathbb{X}, ρ) , a filtration over the real numbers $\text{Filt}(\mathbb{X}, \rho) = (\text{Filt}_\epsilon(\mathbb{X}, \rho))_{\epsilon \in \mathbb{R}}$ is a family of simplicial complexes with vertex set \mathbb{X} such that $\text{Filt}_\epsilon(\mathbb{X}) \subseteq \text{Filt}_{\epsilon'}(\mathbb{X})$ whenever $\epsilon \leq \epsilon'$. For the purposes of this article, we are going to consider only some natural filtrations that are strongly linked to the metric ρ . The *Čech filtration* consists of a family of simplicial complexes $(\check{\text{Cech}}_\epsilon(\mathbb{X}))_{\epsilon \in \mathbb{R}}$ where a set of points $[x_0, \dots, x_k]$ forms a k -simplex of $\check{\text{Cech}}_\epsilon(\mathbb{X})$ if the intersection of the $k+1$ closed balls $\bar{B}_\rho(x_i, \epsilon)$ is non empty. Equivalently, $\check{\text{Cech}}_\epsilon(\mathbb{X})$ is the *nerve* of the cover $\{\bar{B}_\rho(x, \epsilon) : x \in \mathbb{X}\}$. The Čech complex is the most natural way to build a simplicial complex associated to a space, since in favourable cases, it allows to recover its homotopy type as a consequence of the Nerve Theorem (Hatcher, 2002, §4.G). However, the construction of the Čech complex is expensive from a computational point of view, since it requires to check for a large number of intersections. To circumvent this issue, one can instead consider the *Vietoris–Rips filtration* $(\text{Rips}_\epsilon(\mathbb{X}))_{\epsilon \in \mathbb{R}}$. The k -simplices of $\text{Rips}_\epsilon(\mathbb{X})$ are sets $[x_0, \dots, x_k]$ such that $\rho(x_i, x_j) \leq \epsilon$ for all $0 \leq i, j \leq k$. Equivalently, $\text{Rips}_\epsilon(\mathbb{X})$ can be defined as the flag complex of $\check{\text{Cech}}_\epsilon(\mathbb{X})$ (that is, the clique complex of the 1-skeleton of $\check{\text{Cech}}_\epsilon(\mathbb{X})$). If \mathbb{X} is a subset of the Euclidean space \mathbb{R}^D , then one have $\check{\text{Cech}}_\epsilon(\mathbb{X}) \subseteq \text{Rips}_{2\epsilon}(\mathbb{X}) \subseteq \check{\text{Cech}}_{\sqrt{2D/(D+1)}\epsilon}(\mathbb{X})$; see e.g. Theorem 2.5. by de Silva and Ghrist (2007b). In this sense, the Rips complex is a computationally efficient approximation of the Čech complex. Other filtrations involving lower dimensional simplices, such as the *Alpha filtration* by Edelsbrunner et al. (1983), can also be considered in our context.

For any filtration as above, it is clear that the topology of the complexes $\text{Filt}_\epsilon(\mathbb{X})$ will typically change as ϵ increases. This evolution is appropriately captured by considering the homology groups (over a field \mathbf{k}) of the nested family of simplicial complexes. One gets in this way a sequence of vector spaces $(H_\bullet(\text{Filt}_\epsilon(\mathbb{X})))_{\epsilon \in \mathbb{R}}$, where the inclusions $\text{Filt}_\epsilon(\mathbb{X}) \subseteq \text{Filt}_{\epsilon'}(\mathbb{X})$ induce canonical linear maps $H_\bullet(\text{Filt}_\epsilon(\mathbb{X})) \rightarrow H_\bullet(\text{Filt}_{\epsilon'}(\mathbb{X}))$ in homology. Under some conditions, such as finiteness of \mathbb{X} (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005), this sequence can be decomposed as a direct sum of *intervals* $I[\epsilon_b, \epsilon_d]$ defined as

$$0 \xrightarrow{0} \cdots \xrightarrow{0} 0 \xrightarrow{0} \underbrace{\mathbf{k} \xrightarrow{1} \cdots \xrightarrow{1} \mathbf{k}}_{[\epsilon_b, \epsilon_d]} \xrightarrow{0} 0 \xrightarrow{0} \cdots \xrightarrow{0} 0$$

Every interval is determined by the *birth* and *death* parameters ϵ_b and ϵ_d respectively, and it can be interpreted as a *topological feature* of \mathbb{X} with an associated *lifetime* $\epsilon_d - \epsilon_b$ (note that ϵ_d may be infinite, in that case the feature has infinite lifetime). The (multi)set of points (ϵ_b, ϵ_d) is called the *persistence diagram* of (\mathbb{X}, ρ) and is denoted $\text{dgm}(\text{Filt}(\mathbb{X}, \rho))$

(or simply $\text{dgm}(\text{Filt}(\mathbb{X}))$ if ρ is clear from the context). Persistence diagrams are contained in the half (extended) plane above the diagonal $\Delta = \{(x, y) : x = y\}$. For technical reasons, the diagonal Δ is considered as part of every persistence diagram with infinite multiplicity. Chazal et al. (2009, 2016, 2014) proved that, within a more abstract persistent framework, it is possible to extend the definition of persistence diagrams to some cases where the sequence might not be interval-decomposable. In particular, it is shown by Chazal et al. (2014) that if \mathbb{X} is a compact metric space, for every value of ϵ at most a finite number of new topological features appear (even though the vector spaces $(H_\bullet(\text{Filt}_\epsilon(\mathbb{X})))_{\epsilon \in \mathbb{R}}$ may be infinite-dimensional) and hence $\text{dgm}(\text{Filt}(\mathbb{X}))$ is well-defined. Notice also that all the definitions can be extended to filtrations indexed over connected subsets of the real line.

Example 2 (Eyeglasses) *We compute the persistence diagram associated to the Vietoris–Rips filtration of the sample points from Example 1, Figure 1. We compare the results obtained with different distant choices: the Euclidean distance, the k -NN estimator of the inherited Riemannian distance for $k = 4$ and $k = 5$ and the sample Fermat distance for $p = 2.5$ and $p = 3$. We also considered a weighted Vietoris–Rips filtration derived by a DTM-function with parameters $m = 0.01$ and $p = 1$ (see Anai et al. (2019) and Remark 17). The homology of the eyeglasses curve has one generator of H_0 and one generator of H_1 . However, it can be noticed that for either Euclidean and k -NN distance for $k \geq 5$, the persistence diagram displays two salient generators for the first homology group H_1 , which can be attributed to the small reach of the manifold. As it can be seen in Figure 2, smaller values of k fail to capture the geometry of the eyeglasses manifold. A similar situation is presented using the Vietoris–Rips DTM-filtration. Finally, for the Vietoris–Rips filtration using Fermat distance for different choices of p , the diagrams show accurately only one persistent generator for H_1 . On the other hand, the number of noticeable connected components increases with p . This effect is caused by the presence of noisy points in regions of extremely low density, becoming isolated points (or outliers) as p evolves (cf. Remark 16).*

Since in our setup we usually only get an approximation of the metric space under consideration, we will be interested in comparing persistence diagrams built on top of different metric spaces. In this sense, the *bottleneck distance* is a frequently used quantity to measure the difference between two persistence diagrams. Given persistence diagrams dgm_1 and dgm_2 , consider all perfect matchings $M \subseteq \text{dgm}_1 \times \text{dgm}_2$ such that every point of $\text{dgm}_1 \setminus \Delta$ and $\text{dgm}_2 \setminus \Delta$ is paired exactly once in M . Note that points in $\text{dgm}_1 \setminus \Delta$ and $\text{dgm}_2 \setminus \Delta$ are allowed to be paired with points in the diagonal Δ . The bottleneck distance $d_b(\text{dgm}_1, \text{dgm}_2)$ is then defined as the infimum, over all such matchings M as before, of the largest ℓ_∞ -distance between matched pairs. That is,

$$d_b(\text{dgm}_1, \text{dgm}_2) = \inf_M \max_{(x,y) \in M} |x - y|_\infty.$$

The stability theorem (Cohen-Steiner et al., 2007; Chazal et al., 2014) ensures continuity (more precisely, Lipschitz continuity) in the process of computing persistence diagrams for a metric space. This means that small perturbations in the original metric space (in the sense of Gromov–Hausdorff) will translate into an at most proportional perturbation in the

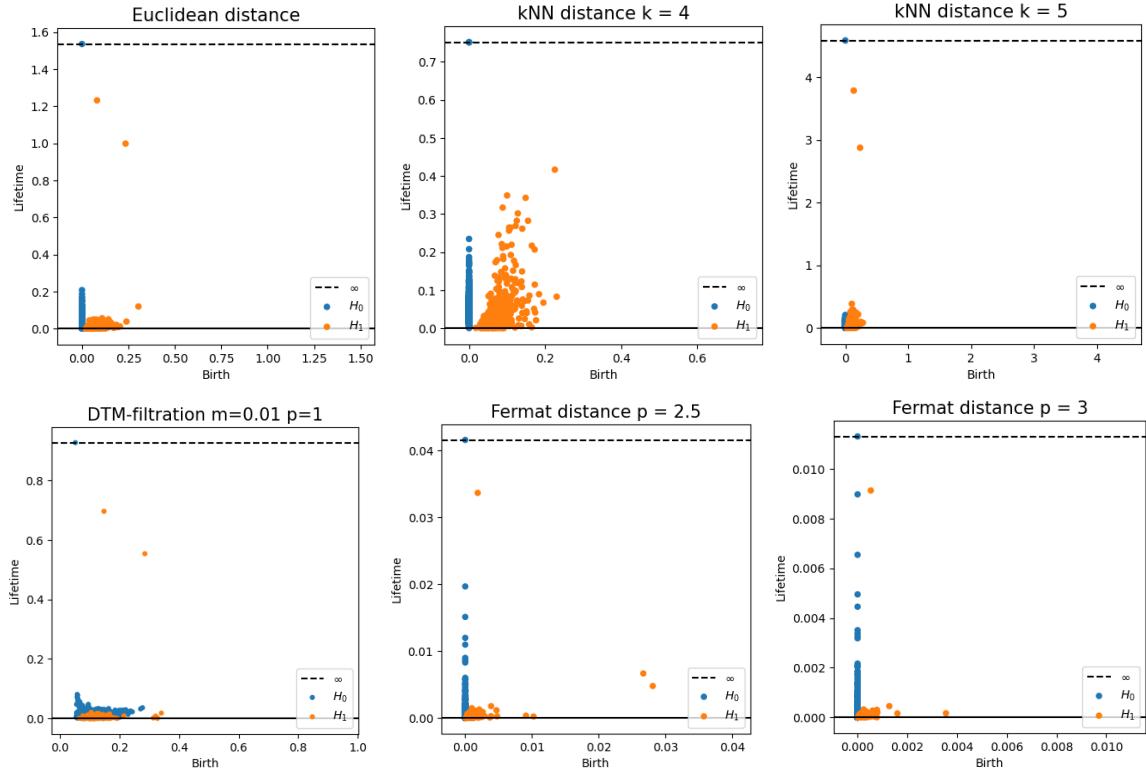


Figure 2: Persistence diagrams (lifetime) associated to the eyeglasses point cloud with noise for different filtrations. Top: Vietoris–Rips filtration with Euclidean distance and k -NN distance for $k = 4$ and $k = 5$. Bottom: Vietoris–Rips DTM-filtration with parameters $m = 0.01$ and $p = 1$ and Vietoris–Rips filtration with Fermat distance for $p = 2.5$ and $p = 3$.

corresponding persistence diagram (in the sense of the bottleneck distance). Formally, it states that for any two precompact metric spaces \mathbb{X} and \mathbb{Y}

$$d_b\left(\text{dgm}(\text{Filt}(\mathbb{X}, \rho_{\mathbb{X}})), \text{dgm}(\text{Filt}(\mathbb{Y}, \rho_{\mathbb{Y}}))\right) \leq 2d_{GH}((\mathbb{X}, \rho_{\mathbb{X}}), (\mathbb{Y}, \rho_{\mathbb{Y}})). \quad (5)$$

This fact is exploited by Fasy et al. (2014); Chazal et al. (2015) to establish the almost sure convergence (in the sense of bottleneck distance) of the persistence diagrams associated to samples of a compact metric space drawn according to a measure satisfying certain hypotheses to the persistence diagram of the space. In these works the distance function of the underlying metric space is assumed to be known, and it is inherited by the sample.

We are able to obtain convergence of persistence diagrams in our context, in which only an estimator of the underlying metric is available. Concretely, given the metric spaces $(\mathcal{M}, d_{f,p})$ and $(\mathbb{X}_n, d_{n,p})$, from the estimation of its Gromov–Hausdorff distance of Theorem 7 and the stability theorem (5) we deduce the following result.

Corollary 10 Let $\varepsilon > 0$ and $\lambda \in ((p-1)/pd, 1/d)$. There exists a constant $\theta > 0$ such that

$$\mathbb{P}\left(d_b(\text{dgm}(\text{Filt}(\mathcal{M}, d_{f,p})), \text{dgm}(\text{Filt}(\mathbb{X}_n, d_{n,p}))) > \varepsilon\right) \leq \exp\left(-\theta n^{(1-\lambda d)\alpha}\right)$$

for n large enough and $\alpha = 1/(d+2p)$.

3.2 Homology Inference

The content of Corollary 10 is that $\text{dgm}(\text{Filt}(\mathbb{X}_n, d_{n,p}))$ is (asymptotically) a good estimator of $\text{dgm}(\text{Filt}(\mathcal{M}, d_{f,p}))$. On the other hand, if we were to employ the Euclidean distance $|\cdot|$, it follows from the results by Chazal et al. (2015) that the sample persistence diagrams $\text{dgm}(\text{Filt}(\mathbb{X}_n, |\cdot|))$ converge to $\text{dgm}(\text{Filt}(\mathcal{M}, |\cdot|))$ under reasonable hypotheses. We are therefore interested in comparing for how long we may expect to read the correct homology of \mathcal{M} in each of the diagrams $\text{dgm}(\text{Filt}(\mathcal{M}, d_{n,p}))$ and $\text{dgm}(\text{Filt}(\mathcal{M}, |\cdot|))$ in terms of two natural geometric measures associated to the manifold, namely, the reach and the convexity radius (see Hausmann, 1995; Latschev, 2001; Niyogi et al., 2008; Chazal and Lieutier, 2008). In this section we show that the homology of $(\mathcal{M}, d_{f,p})$ can be recovered correctly from its persistence diagram up to the convexity radius $\text{conv}(\mathcal{M}, d_{f,p})$, whereas for $(\mathcal{M}, |\cdot|)$ this is guaranteed only up to its reach $\tau_{\mathcal{M}}$. Notice that the reach of a submanifold of an Euclidean space depends strongly on the particular embedding, whereas the convexity radius is an intrinsic quantity linked to the geometry of the manifold. There are simple examples of manifolds in which this distinction is relevant to correctly recover its homology from a sample (see Examples 1 and 3).

Recall that given $\mathbb{X} \subseteq \mathbb{R}^D$ a closed subset, the *medial axis* $\text{Med}(\mathbb{X})$ of \mathbb{X} is defined as

$$\text{Med}(\mathbb{X}) := \{y \in \mathbb{R}^D : d_E(y, \mathbb{X}) = |p - y| \text{ for at least two different points } p \in \mathbb{R}^D\},$$

where $d_E(y, \mathbb{X}) = \inf_{x \in \mathbb{X}} |y - x|$. The *reach* $\tau_{\mathbb{X}}$ of \mathbb{X} , first introduced by Federer (1959), is the minimum distance from \mathbb{X} to $\text{Med}(\mathbb{X})$, that is,

$$\tau_{\mathbb{X}} := \inf_{x \in \mathbb{X}} d_E(x, \text{Med}(\mathbb{X})).$$

Given a Riemannian manifold (\mathcal{N}, g) , we will say that a subset $S \subseteq \mathcal{N}$ is *geodesically convex* if for every two points in S , there is a unique geodesic segment that connects them and it is completely contained in S . The *convexity radius* $\text{conv}(\mathcal{N}, x)$ at a point $x \in \mathcal{N}$ is the supremum over those $r > 0$ for which the (geodesic) ball $B(x, r)$ is geodesically convex. The convexity radius $\text{conv}(\mathcal{N})$ of the manifold \mathcal{N} is defined as

$$\text{conv}(\mathcal{N}) := \inf_{x \in \mathcal{N}} \text{conv}(\mathcal{N}, x).$$

Proposition 11 Let \mathcal{M} be a compact submanifold of \mathbb{R}^D . Then, we have the following homotopy equivalences:

- $\check{\text{Cech}}_{\epsilon}(\mathcal{M}, |\cdot|) \simeq \mathcal{M}$ for $\epsilon < \tau_{\mathcal{M}}$ and $\text{Rips}_{\epsilon}(\mathcal{M}, |\cdot|) \simeq \mathcal{M}$ for $\epsilon < 2\sqrt{\frac{D+1}{2D}}\tau_{\mathcal{M}}$, and both bounds are optimal, in the sense that there exist examples for which the homotopy equivalence does not hold for larger values of ϵ .

- $\check{\text{C}}\text{ech}_\epsilon(\mathcal{M}, d_{f,p}) \simeq \mathcal{M}$ and $\text{Rips}_\epsilon(\mathcal{M}, d_{f,p}) \simeq \mathcal{M}$ for $\epsilon < \text{conv}(\mathcal{M}, d_{f,p})$.

Moreover, if $d_{f,p}$ coincides up to a constant with $d_{\mathcal{M}}$ (i.e. f is uniform), we have the estimate

$$\text{conv}(\mathcal{M}, d_{f,p}) = \text{Vol}(\mathcal{M}, d_{\mathcal{M}})^{(p-1)/d} \text{conv}(\mathcal{M}, d_{\mathcal{M}}) \geq \text{Vol}(\mathcal{M}, d_{\mathcal{M}})^{(p-1)/d} \frac{\pi}{2} \tau_{\mathcal{M}}.$$

Proof The fact that $\check{\text{C}}\text{ech}_\epsilon(\mathcal{M}, |\cdot|)$ is homotopy equivalent to \mathcal{M} for $\epsilon < \tau_{\mathcal{M}}$ is an immediate consequence of the Nerve Theorem. The same result implies that $\check{\text{C}}\text{ech}_\epsilon(\mathcal{M}, d_{f,p}) \simeq \mathcal{M}$ for $\epsilon < \text{conv}(\mathcal{M}, d_{f,p})$, since geodesically convex sets are always contractible and the intersection of geodesically convex sets is again geodesically convex. Regarding the Vietoris–Rips filtration, the fact that the simplicial complex $\text{Rips}_\epsilon(\mathcal{M}, |\cdot|)$ is homotopy equivalent to \mathcal{M} for $\epsilon < 2\sqrt{\frac{D+1}{2D}}\tau_{\mathcal{M}}$ can be deduced from (Kim et al., 2020, Theorem 20). Finally, since $d_{f,p}$ is a Riemannian distance on \mathcal{M} , by Hausmann (1995) there is an explicit homotopy equivalence $\text{Rips}_\epsilon(\mathcal{M}, d_{f,p}) \simeq \mathcal{M}$ for $\epsilon < \text{conv}(\mathcal{M}, d_{f,p})$ (see also Latschev, 2001).

The optimality of the bound $\epsilon < \tau_{\mathcal{M}}$ for $\check{\text{C}}\text{ech}_\epsilon(\mathcal{M}, |\cdot|)$ is clear (think of a unit sphere in \mathbb{R}^D), and indeed, typically the topology of $\check{\text{C}}\text{ech}_\epsilon(\mathcal{M}, |\cdot|)$ changes when ϵ attains $\tau_{\mathcal{M}}$. A critical example for the Vietoris–Rips complex is the standard 1-dimensional circle \mathbb{S}^1 , and it can be derived from the main result of Adamaszek and Adams (2017), similarly as in (Kim et al., 2020, Example 24).

The last assertion in the statement follows directly from the inequalities

$$\text{conv}(\mathcal{M}, d_{\mathcal{M}}) \geq \min \left\{ \frac{\pi}{2\sqrt{\sup K}}, \frac{1}{2} \text{inj}(\mathcal{M}, d_{\mathcal{M}}) \right\}$$

(see Cheeger and Ebin, 1975, §5.14) and

$$\text{inj}(\mathcal{M}, d_{\mathcal{M}}) \geq \pi \tau_{\mathcal{M}}, \quad K \leq \frac{1}{\tau_{\mathcal{M}}^2}$$

(see Aamari et al., 2019, Proposition A.1). Here $\text{inj}(\mathcal{M}, d_{\mathcal{M}})$ is the injectivity radius of \mathcal{M} and K is the sectional curvature. \blacksquare

Example 3 Consider a planar ellipse $E_{R,\varepsilon}$ with minor axis of length ε and major axis of length $R \geq \varepsilon$. By letting $R \rightarrow +\infty$ and/or $\varepsilon \rightarrow 0$, we see that the convexity radius of a closed submanifold of \mathbb{R}^2 can be arbitrarily large while its reach can be arbitrarily small. A similar example can be constructed in \mathbb{R}^D , being \mathcal{M} a d -dimensional ellipsoid for any $d < D$. The same phenomenon can be achieved by constructing different eyeglasses curves with arbitrarily large length and constant reach, Figure 3. Its population persistence diagrams differ as predicted by Theorem 11. The persistence diagram computed with the Euclidean distance captures the right homology only for ϵ less than the reach. In contrast, for the Fermat distance the correct homology is captured for radii as large as (a multiple of) the convexity radius, which can be made large enough by enlarging the bridge between the glasses.

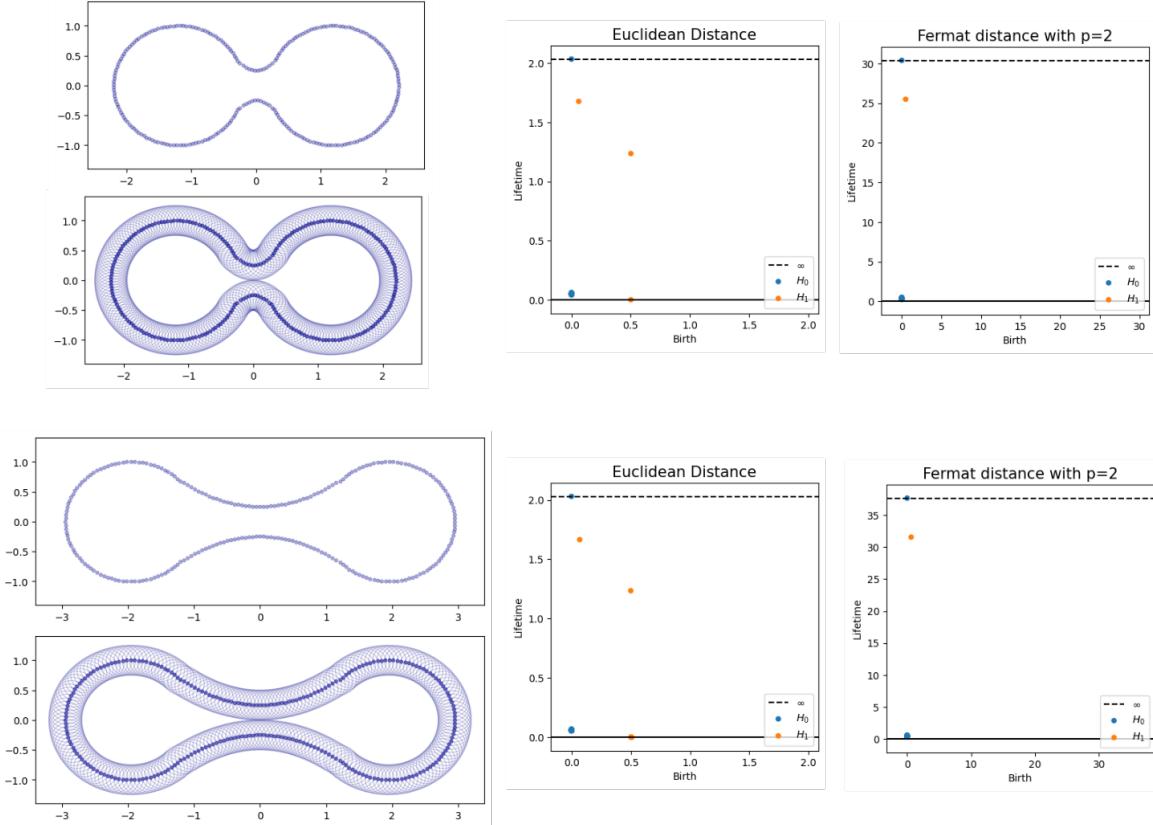


Figure 3: Left: Eyeglasses curves, uniformly sampled (250 points). In both cases, the reach is 0.5. Below each curve, we plot a thickening of the samples with Euclidean balls of radius slightly greater than the reach. Right: Persistence diagrams (lifetime) associated to the Vietoris–Rips filtration for both the Euclidean distance and the re-scaled Fermat distance $d_{n,p}$ with $p = 2$. While H_0 is correctly estimated in both cases by reading the persistence diagrams, the ones computed with the Euclidean distance displays two salient generators for the first homology group H_1 , inaccurately suggesting two cycles. The second cycle's birth is at the level of twice the reach. For the (re-scaled) Fermat distance, the diagrams shows correctly only one persistent generator for H_1 .

3.3 Robustness to Outliers

Persistence diagrams are highly sensitive to outliers (see Bendich et al., 2011; Chazal et al., 2011; Buchet et al., 2016; Anai et al., 2019). We will see that the computation of persistence homology using Fermat distance is robust to the presence of outliers for positive degree. Concretely, given a sample $\mathbb{X}_n \subseteq \mathcal{M}$ and $Y \subseteq \mathbb{R}^D \setminus \mathcal{M}$ a finite set of points in the complement of \mathcal{M} in the ambient Euclidean space — the *outliers* — we prove that $\text{dgm}_k(\text{Rips}(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p}))$ coincides with $\text{dgm}_k(\text{Rips}(\mathbb{X}_n, d_{\mathbb{X}_n, p}))$ for $k > 0$ up to some reasonable filtration parameter. First we need a definition.

Definition 12 Given a finite set of points $S \subseteq \mathbb{R}^D$, define the minimal spacing of S as

$$\kappa(S) = \min_{x \in S} d_E(x, S \setminus \{x\}),$$

where d_E denotes the Euclidean distance between sets.

Proposition 13 Let $\delta = \min\{\kappa(Y), d_E(\mathbb{X}_n, Y)\}$ and $p > 1$. Then, for every $\epsilon < \delta^p$

$$\text{Rips}_\epsilon(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p}) = \text{Rips}_\epsilon(\mathbb{X}_n, d_{\mathbb{X}_n, p}) \cup Y.$$

In particular, for all $k > 0$

$$\text{dgm}_k(\text{Rips}_{<\delta^p}(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})) = \text{dgm}_k(\text{Rips}_{<\delta^p}(\mathbb{X}_n, d_{\mathbb{X}_n, p})),$$

where $\text{Rips}_{<\delta^p}(\mathbb{X}, \rho_{\mathbb{X}})$ stands for $(\text{Rips}_\epsilon(\mathbb{X}, \rho_{\mathbb{X}}))_{\epsilon < \delta^p}$, i.e., the Rips filtration up to parameter δ^p of a metric space $(\mathbb{X}, \rho_{\mathbb{X}})$.

Proof Let us estimate the distance between two given points in $\mathbb{X}_n \cup Y$ with respect to $d_{\mathbb{X}_n \cup Y, p}$ in terms of δ and $d_{\mathbb{X}_n, p}$.

If $x \in \mathbb{X}_n$ and $y \in Y$,

$$d_{\mathbb{X}_n \cup Y, p}(x, y) \geq d_{\mathbb{X}_n \cup Y, p}(\mathbb{X}_n, Y) = d_E(\mathbb{X}_n, Y)^p \geq \delta^p.$$

If $y, y' \in Y$,

$$d_{\mathbb{X}_n \cup Y, p}(y, y') \geq d_{\mathbb{X}_n \cup Y, p}(y, Y \setminus \{y\}) \geq \delta^p.$$

For the second inequality, notice that if $\tilde{y} \in Y$ is such that $d_{\mathbb{X}_n \cup Y, p}(y, Y \setminus \{y\}) = d_{\mathbb{X}_n \cup Y, p}(y, \tilde{y}) = \text{len}(\gamma)$, the geodesic γ between y and \tilde{y} either involves only points from Y or there exist some point $x \in \mathbb{X}_n$ in γ . In the first case $d_{\mathbb{X}_n \cup Y, p}(y, \tilde{y}) \geq \kappa(Y)^p$ whereas in the second case $d_{\mathbb{X}_n \cup Y, p}(y, \tilde{y}) \geq 2d_E(\mathbb{X}_n, Y)^p$.

Given $x, x' \in \mathbb{X}_n$, let γ be a minimal path between x, x' , so that $d_{\mathbb{X}_n \cup Y, p}(x, x') = \text{len}(\gamma)$. If $d_{\mathbb{X}_n \cup Y, p}(x, x') < \epsilon$, then γ only involves points in \mathbb{X}_n since otherwise $\epsilon \geq \text{len}(\gamma) \geq 2d_E(\mathbb{X}_n, Y) \geq 2\delta^p$, which is a contradiction. Hence, $d_{\mathbb{X}_n \cup Y, p}(x, x') = d_{\mathbb{X}_n, p}(x, x')$. ■

We define now a geometric notion of outliers. Recall that given $\mathbb{X}_n \subseteq \mathbb{R}^D$, the ε -graph $G_\varepsilon(\mathbb{X}_n)$ is the undirected graph with the points of \mathbb{X}_n as vertices and an edge connecting x_i and $x_j \in \mathbb{X}_n$ whenever $|x_i - x_j| < \varepsilon$.

Definition 14 Let $\mathbb{X}_n \subseteq \mathcal{M}$ be a sample of $\mathcal{M} \subseteq \mathbb{R}^D$ and $Y \subseteq \mathbb{R}^D \setminus \mathcal{M}$ be a finite set of points. Let $\varepsilon_* := \min\{\varepsilon > 0 : G_\varepsilon(\mathbb{X}_n) \text{ is connected}\}$ and $\delta = \min\{\kappa(Y), d_E(\mathbb{X}_n, Y)\}$. We say that Y are (geometric) outliers if $\delta > \varepsilon_*$.

We show next that for this notion of outliers, the upper bound on the parameter for the Rips filtration of Proposition 13 is not restrictive for sufficiently large p . Indeed, let $\text{diam}_p(\mathbb{X}_n)$ be the diameter of $(\mathbb{X}_n, d_{\mathbb{X}_n, p})$. Note that for every $\epsilon \geq \text{diam}_p(\mathbb{X}_n)$ the simplicial complex $\text{Rips}_\epsilon(\mathbb{X}_n, d_{\mathbb{X}_n, p})$ equals the standard $(n-1)$ -simplex Δ^{n-1} , with trivial topology (and hence persistence diagrams are not interesting for scales larger than this threshold). The next result states that provided that p is large enough, the persistence diagrams of $(\mathbb{X}_n, d_{\mathbb{X}_n, p})$ and $(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})$ coincide up to the filtration parameter $\text{diam}_p(\mathbb{X}_n)$.

Corollary 15 Given \mathbb{X}_n a sample of \mathcal{M} and $Y \subseteq \mathbb{R}^D$ a finite set of outliers, then for all $k > 0$

$$\mathrm{dgm}_k(\mathrm{Rips}_{<\mathrm{diam}_p(\mathbb{X}_n)}(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})) = \mathrm{dgm}_k(\mathrm{Rips}_{<\mathrm{diam}_p(\mathbb{X}_n)}(\mathbb{X}_n, d_{\mathbb{X}_n, p})).$$

for $p > C \log(n)$ with $C = \log(\delta/\epsilon_*)^{-1}$.

Proof There is an upper bound $\mathrm{diam}_p(\mathbb{X}_n) \leq n\varepsilon_*^p$. Since Y are outliers, $\varepsilon_* < \delta$. For $p > C \log(n)$, $\left(\frac{\delta}{\varepsilon_*}\right)^p > n$ and consequently, $\mathrm{diam}_p(\mathbb{X}_n) < \delta^p$. The result now follows from Proposition 13. \blacksquare

Remark 16 In general, the persistence diagram of $(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})$ for degree $k = 0$ does not coincide with the diagram of the metric space without outliers $(\mathbb{X}_n, d_{\mathbb{X}_n, p})$. However, if Y is a set of geometric outliers, it is related to the corresponding persistence diagrams of \mathbb{X}_n and Y through the following formula:

$$\mathrm{dgm}_0(\mathrm{Rips}(\mathbb{X}_n \cup Y, d_{\mathbb{X}_n \cup Y, p})) = \mathrm{dgm}_0^{<\infty}(\mathrm{Rips}(\mathbb{X}_n, d_{\mathbb{X}_n, p})) \cup \mathrm{dgm}_0(\mathrm{Rips}(Q, d_Q)).$$

Here, $\mathrm{dgm}^{<\infty}$ denotes the bounded persistence intervals and $Q = (Y \cup \mathbb{X}_n)/\mathbb{X}_n$ is the quotient metric space endowed with the induced metric d_Q .

Remark 17 (DTM) Filtrations classically used for the computation of persistent homology of Euclidean point clouds, such as the Čech or Vietoris–Rips filtrations, are very sensitive to the presence of outliers. That is, Čech (or Vietoris–Rips) filtrations computed on top of \mathbb{X}_n and $\mathbb{X}_n \cup Y$ might be very different (its interleaving distance depends on $d_H(\mathbb{X}_n, \mathbb{X}_n \cup Y)$, see e.g. Chazal et al. (2014)). To overcome this limitation, Anai et al. (2019) introduced weighted filtrations based on the notion of distance to measure (DTM). Given μ the empirical measure of $\mathbb{X}_n \subseteq \mathbb{R}^D$ and $m \in [0, 1)$ a parameter, the DTM-function over \mathbb{R}^D is defined as $d_{\mu, m}(x) := \sqrt{\frac{1}{m} \int_0^m \delta_{\mu, t}^2(x) dt}$, where $\delta_{\mu, t}(x) = \inf\{r \geq 0 : \mu(\bar{B}(x, r)) > t\}$ and $\bar{B}(x, r)$ denotes the closed Euclidean ball with center x and radius r . Given a parameter $p > 1$, the weighted ball $B_{d_{\mu, m}}(x, \epsilon)$ with center $x \in \mathbb{X}_n$ and radius $\epsilon \geq d_{\mu, m}(x)$ is the Euclidean ball $B(x, r_x(\epsilon))$ with radius $r_x(\epsilon) = (\epsilon^p - d_{\mu, m}^p(x))^{1/p}$ (if $\epsilon < d_{\mu, m}(x)$, it is empty). The Čech DTM-filtration $(V_{m, p}^{\mathrm{DTM}}(\mathbb{X}_n))_{\epsilon > 0}$ with parameters (m, p) is the weighted Čech filtration constructed as the nerve of the cover $\{B_{d_{\mu, m}}(x, \epsilon) : x \in \mathbb{X}_n\}$ for every $\epsilon > 0$. A DTM-based version of a weighted Vietoris–Rips filtration can also be derived.

DTM-filtrations of Euclidean point clouds produce filtrations (and hence, persistence diagrams) less sensitive to outliers, given that the (interleaving) distance between $V_{m, p}^{\mathrm{DTM}}(\mathbb{X}_n)$ and $V_{m, p}^{\mathrm{DTM}}(\mathbb{X}_n \cup Y)$ is upper bounded not only in terms of $d_H(\mathbb{X}_n, \mathbb{X}_n \cup Y)$ but also in terms of the Wasserstein distance between the measures $\mu_{\mathbb{X}_n}$ and $\mu_{\mathbb{X}_n \cup Y}$. However, if \mathbb{X}_n is a sample of a manifold \mathcal{M} , these filtrations are still very sensitive to the particular embedding of the manifold in \mathbb{R}^D . This is consequence of the dependence of the DTM-function on the ambient space (see Example 4). Its (lack of) dependence on non-intrinsic properties has been investigated thereafter. In this direction, a generalization of DTM-filtrations for general metric spaces (\mathbb{X}, ρ) is considered in (Buchet et al., 2016).

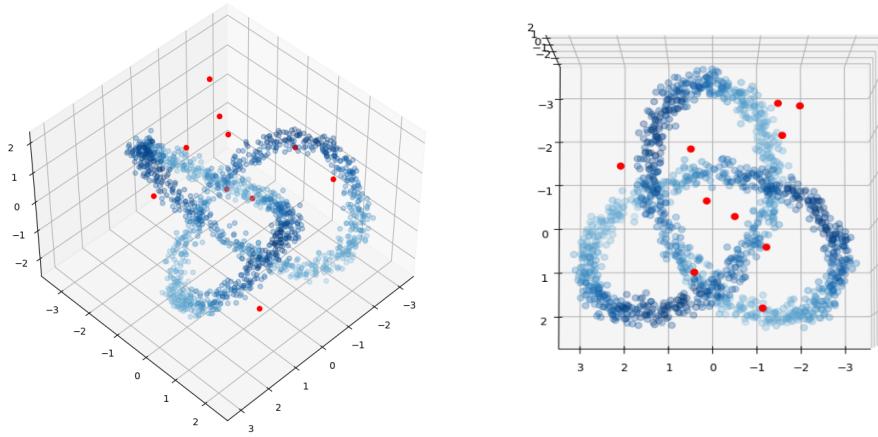


Figure 4: A (noisy) sample of 1500 points from the trefoil knot with outliers (red).

Example 4 (Trefoil) Consider the embedding of a topological circle \mathbb{S}^1 in \mathbb{R}^3 given by the trefoil knot. In particular, it is homeomorphic to \mathbb{S}^1 and its homology has just one generator in H_0 (one connected component) and one generator in H_1 (one 1-dimensional cycle). Given a (noisy) sample of 1500 points from the trefoil knot with 10 outliers, Figure 4, we compute its persistence diagram for different choices of filtrations and compare them with the case without the outliers, Figure 5. For the Vietoris–Rips filtration using Euclidean distance, the small reach of the embedding produces a persistence diagram with four persistent generators for H_1 in both cases, with and without outliers (cf. Example 3). If we use k -NN distances, the presence of outliers affects the accuracy of the topological features captured in the persistence diagram, which presents four salient generators for H_1 instead of the single generator recovered from the sample without outliers. For the Vietoris–Rips DTM-filtration, we observe that the diagrams are comparable both in absence and presence of outliers. However, the dependence of the embedding of the construction is reflected in the incorrect number of generators for H_1 with long persistence. Finally, the persistence diagram computed from the Vietoris–Rips filtration using Fermat distance remains unaffected in presence of outliers for degree 1 (Corollary 15), and it shows correctly a single salient generator of H_1 . For degree 0, the diagram is related to the diagram of the sample without the outliers and the diagram of the outliers themselves (cf. Remark 16).

3.4 Computational Complexity

Our proposed pipeline for the computation of Fermat-based persistent homology consists of the precomputation of Fermat distance in the input sample \mathbb{X}_n , followed by the computation of persistent homology from the metric space $(\mathbb{X}_n, d_{\mathbb{X}_n, p})$ described by the distance matrix.

The computation of the matrix of pairwise sample Fermat distances between points in \mathbb{X}_n has complexity $\mathcal{O}(n^3)$. However, it can be reduced to $\mathcal{O}(n^2 \log^2 n)$ with high probability by restricting the computation of shortest paths to the k -NN graph on top of \mathbb{X}_n with $k = O(\log n)$ (see Section 2.3 in Groisman et al. (2022), Little et al. (2022); Chu et al. (2020)).

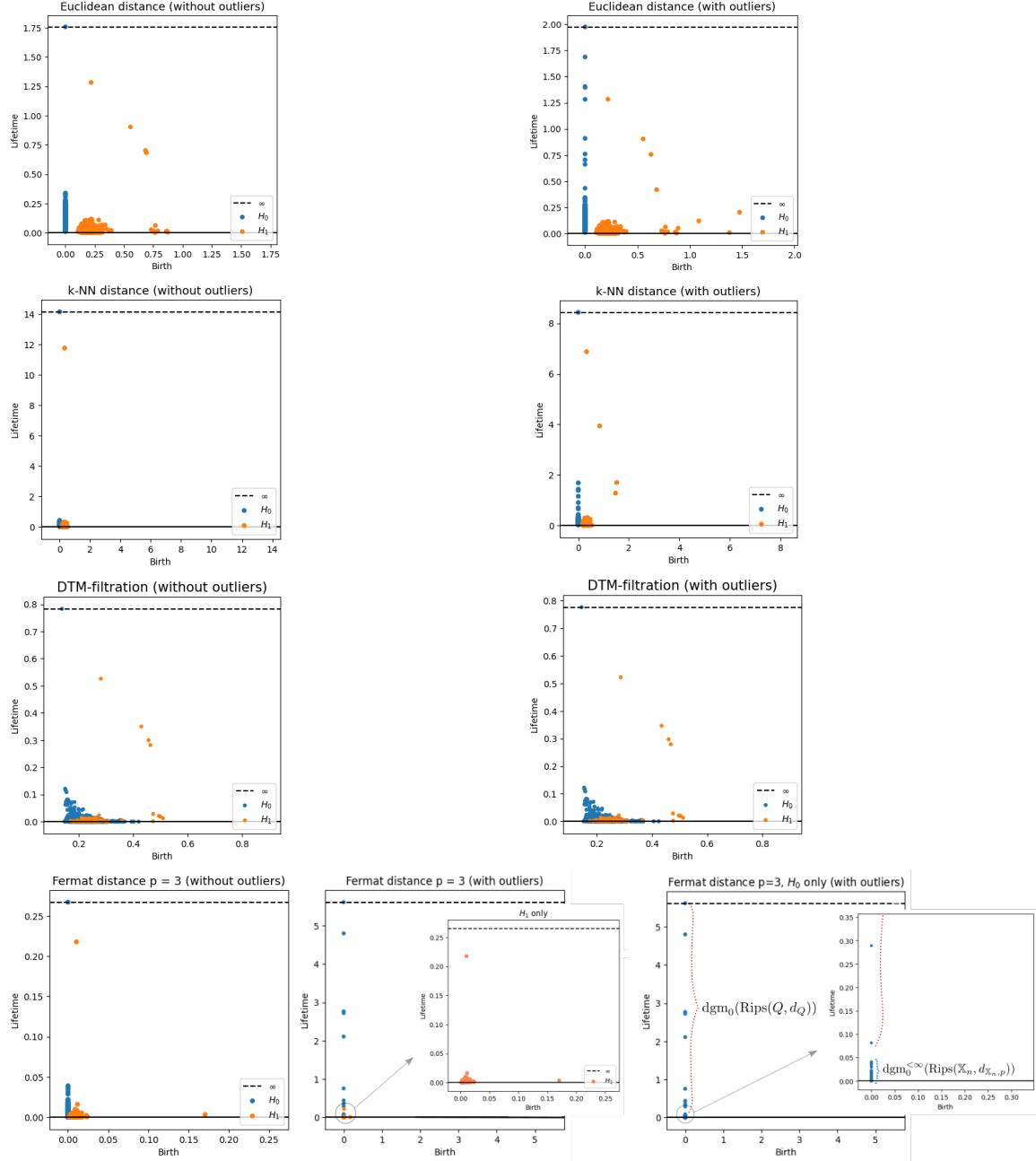


Figure 5: Persistence diagrams associated to the Vietoris-Rips filtration of the sample of the trefoil knot using Euclidean distance, k -NN distance with $k = 10$, DTM weight and Fermat distance with $p = 3$ of the sample without outliers \mathbb{X}_n (left) and the sample with outliers $\mathbb{X}_n \cup Y$ (right) respectively. When Fermat distance is used, the persistence diagram of $\mathbb{X}_n \cup Y$ for degree 1 equals the diagram of \mathbb{X}_n (without outliers). For degree 0, it decomposes as the union of the subdiagram of finite intervals of \mathbb{X}_n , $dgm_0^{<\infty}(\text{Rips}(\mathbb{X}_n, d_{\mathbb{X}_n, p}))$, and the diagram $dgm_0(\text{Rips}(Q, d_Q))$ of the quotient space $Q = (Y \cup \mathbb{X}_n)/\mathbb{X}_n$.

On the other hand, the *standard algorithm* used to compute persistent homology was first introduced by Edelsbrunner et al. (2002) and it is based on the Gaussian reduction of the boundary matrix. Persistent homology for degree up to k depends on the $(k+1)$ -skeleton of the filtration and the worst case computational complexity is cubical in the number N of simplices of dimension at most $k+1$ (Morozov, 2005; Otter et al., 2017). An alternative algorithm for the reduction of the boundary matrix, introduced by Milosavljević et al. (2011), has complexity $O(N^\omega)$, with ω the matrix multiplication coefficient. At present, the best bound for ω is 2.376 (Coppersmith and Winograd, 1987).

In practice, computation of persistent homology has lower complexity. For Vietoris–Rips filtrations, the worst case complexity is for k -dimensional persistent homology is $O\left(\binom{n}{k+2}^3\right) = O(n^{3(k+2)})$ with n the number of vertices of \mathbb{X}_n . However, Giunti et al. (2022) proved that, for instance, the average complexity for the reduction of the boundary matrix of degree 1 is upper bounded by $O(n^5 \log^2(n))$. Moreover, they showed that this upper bound seems to be not tight, since experimental simulations show that the average cost of the reduction of the 1-boundary matrix follows a curve of around $O(n^{3.73})$.

Overall, our proposed pipeline based on the precomputation of pairwise Fermat distance in \mathbb{X}_n does not increase the complexity of the total persistent homology computation.

4. Applications to Signal Analysis

In this section we present a method for change-point detection and pattern recognition in time series through the analysis of topological features (see also Maletić et al. (2016); Perea (2019); Perea and Harer (2015)). This method is illustrated by a series of experiments in both synthetic and real data. In the experiments, the use of Fermat distance (as opposed to Euclidean distance) is observed to lead to more robust inference of the topology of the underlying space. We remark that in these examples the data does not necessarily verify the i.i.d. assumption.

Fermat and k -NN distances are computed using the library `Fermat` (Aristas, 2018), while `Ripser` by Bauer (2021) is employed for the computation of persistence diagrams associated to Vietoris–Rips filtrations. All the computations are over the field $\mathbf{k} = \mathbb{Z}_2$. The code for all the examples and experiments can be found in the repository by Fernandez (2021a).

4.1 Topological Analysis of Time Series.

Time-delay embeddings of scalar time-series data is a well-known technique to recover the underlying dynamics of a system. Takens (1981) theorem gives conditions under which a smooth attractor can be reconstructed from a generic observable function, with dimensional bounds related to those of the Whitney Embedding Theorem. It implies in particular that if $X(t)$ is a real valued signal (which is assumed to be one of the coordinates of a flow given by a system of differential equations), then the *delay coordinate map*

$$t \mapsto (X(t), X(t + \tau), X(t + 2\tau), \dots, X(t + (D - 1)\tau))$$

is an embedding of an orbit. Here D is the embedding dimension and τ is the time delay. From a theoretical point of view, D is the number of variables of the original system. However, in practice the underlying equations describing the dynamical system are not

available. Thus, dynamics are often analyzed by studying the topology of their *attractors*; i.e., invariant subsets of the phase space towards which the system tends to evolve (Birman and Williams, 1983; Smale, 1967; Gilmore and Lefranc, 2002). If the attractor is a smooth manifold \mathcal{M} of dimension d , under certain conditions Takens' theorem implies that the delay embedding of the signal with $D \geq 2d + 1$ is diffeomorphic to \mathcal{M} .

We describe now an approach — based on intrinsic persistence diagrams — to study geometry of attractors and pattern recognition in time series by means of the analysis of the time evolving topological organization of the embedded flow. Let (x_1, x_2, \dots, x_n) be a time series, i.e. a finite sample of a signal $X : [0, T] \rightarrow \mathbb{R}$ such that for evenly spaced points $0 = t_1 < t_2 < \dots < t_n = T$, $x_i = X(t_i)$ for all $1 \leq i \leq n$. Given D and τ , compute the delay embedding of the time series

$$\mathbb{X}_n = \{(x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(D-1)\tau}) : 1 \leq i \leq n - (D-1)\tau\} \subseteq \mathbb{R}^D.$$

Then, for $p > 1$, endow \mathbb{X}_n with a metric space structure induced by the sample Fermat distance $d_{\mathbb{X}_n, p}$. The persistence diagram of the delay embedding $(\mathbb{X}_n, d_{\mathbb{X}_n, p})$ quantifies information about the homology of the attractor associated to the underlying dynamical system.

Example 5 (Reconstruction of Lorenz attractor) *The parameters associated to the delay coordinate reconstruction for a time series can be determined following some heuristics (e.g. false nearest neighbors to determine the embedding dimension (Kennel et al., 1992)). However, in case of noisy data, the embedding dimension is often over-estimated and it may have a great impact on the phase space reconstruction. Indeed, in high dimensional spaces, any two points of a typical large set are at similar Euclidean distance (Aggarwal et al., 2001). This phenomenon is part of what is known as the curse of dimensionality. For this reason, the choice of an intrinsic distance is crucial to recover the right topological features of a space embedded in high dimension.*

Consider the strange attractor associated to the Lorenz (1963) system

$$\begin{cases} \dot{x} = \sigma(y - x), \\ \dot{y} = x(\rho - z) - y, \\ \dot{z} = xy - \beta z \end{cases} \quad (6)$$

when $(\sigma, \rho, \beta) = (10, 28, 8/3)$.

In Figure 6 we take a numerical integration $\varphi(t, v_0)$ of (6) with $dt = 0.01$, satisfying the initial condition $\varphi(0, v_0) = v_0$ with $v_0 = (1, 1, 1)$. We inspect the time series corresponding to the x -coordinate with additive Gaussian noise with variance 0.1, and recover topological information of the attractor from the delay embedding (see also Maletić et al. (2016)). Notice that in this case, although the number of variables in the underlying system is 3, the dimension of the attractor is $d = 2$ so the embedding dimension estimated by Takens' theorem is greater than or equal to 5.

The persistence diagram of the delay embedding reconstruction is computed with time delay $\tau = 10$ and embedding dimensions $D = 3, 4$ and 5 , Figure 6. Here, a uniform down-sampling from the original point cloud of ~ 10000 points is computed, to obtain a new point cloud of ~ 3400 points.

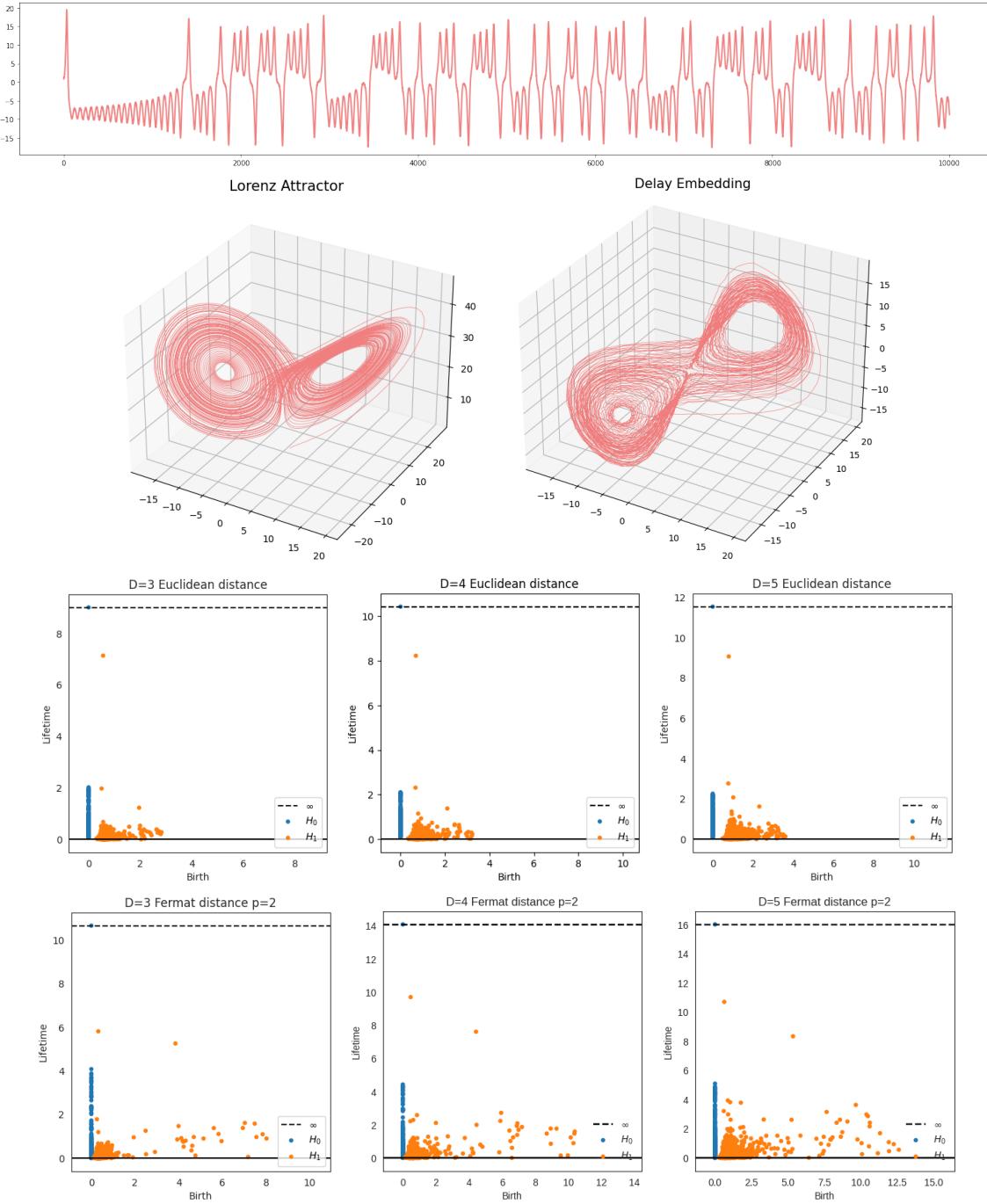


Figure 6: From top to bottom: The x -coordinate time series with Gaussian noise (variance = 0.1) of the Lorenz attractor. The original trajectory and the delay embedding of the noisy x -coordinate time series with $D = 3$ and $\tau = 10$. Persistence diagrams associated to the delay embedding computed with Euclidean and Fermat distances for embedding dimension $D = 3$, $D = 4$ and $D = 5$ and time delay $\tau = 10$.

The Lorenz attractor is homotopy equivalent to the eight-space with two holes corresponding to the equilibrium points that the trajectory never reaches. As Figure 6 reveals, the use of Fermat distance leads to robustly capturing the intrinsic two 1-cycles for the different embedding dimensions, while this is not the case for the Euclidean distance.

Example 6 (Periodicity) A periodic dynamic within a noisy system might be robustly captured using time-delay embeddings. Indeed, embeddings of periodic signals have the topology of a cycle. However, the general success of the reconstruction of the intrinsic cyclic geometry is highly dependent on the choice of the delay parameter τ (and the embedding dimension D). In practice, classic heuristics based on time-delayed mutual information (Fraser and Swinney, 1986) and false nearest neighbors (Kennel et al., 1992) are used, but they present high sensitiveness to noise. We show that the use of Fermat distance when recovering the intrinsic geometry of delay embeddings has stability properties with respect to the choice of τ .

Consider the function $f(t) = \cos(t) + \cos(3t)$ with additive Gaussian noise of variance 0.4. For a sample of 2000 points of the noisy signal in consideration at the interval $[0, 100]$, the classic heuristic estimations of the optimal parameters outputs $\tau = 28$ and $D = 8$ (here, the computations are preformed with the package *Time Series* from the software *Giotto-tda* by Tauzin et al. (2020)). However, the associated time-delay embedding presents low reach value and, hence, it is still hard to capture its homology with standard methods (see Figure 7).

In general dynamics, the effect of the choice of τ is reflected in changes in the embedding of the associated attractor in the ambient space. Although Takens' theorem theoretically establishes diffeomorphic embeddings for different choices of τ , in practice the accuracy of the reconstruction of the underlying manifold usually depends on the choice of τ . Crucially, persistence diagrams computed using Fermat distance are less dependent of extrinsic properties and hence, highly appropriate for the estimation of topological properties of the attractor (that are, indeed, independent of the embedding). To illustrate the stability properties with respect to the choice of the delay parameter, we computed the delay embedding of the noisy periodic signal of Figure 7 in \mathbb{R}^8 for a range of values of τ . We observe that, while the features displayed on the diagrams computed using Euclidean distance change with the embedding, the ones computed using Fermat distance are consistent: they all display a single generator for H_1 (Figure 8). Here, p was set equal to 6, but similar results can be obtained for a range of values of p .

In order to identify changes in patterns of time series, we investigate the topological evolution in time of the delay embedding. For every sample time $t_j \in [0, T]$ ($1 \leq j \leq n - (D - 1)\tau$), consider the delay embedding \mathbb{X}_j of the restriction of the time series up to time t_j , with the metric structure inherited from $(\mathbb{X}_n, d_{\mathbb{X}_n, p})$. That is,

$$\mathbb{X}_j := \{(x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(D-1)\tau}) : 1 \leq i \leq j\} \subseteq \mathbb{X}_n.$$

If $\mathcal{M}[0, t]$ is the delay embedding of the restricted signal $X|_{[0, t]}$, the time evolving series of diagrams $\{\text{dgm}(\text{Rips}(\mathbb{X}_i)) : 1 \leq i \leq n - (D - 1)\tau\}$ is a sample of an approximation of the curve

$$t \mapsto \text{dgm}(\text{Rips}(\mathcal{M}[0, t])), \quad (7)$$

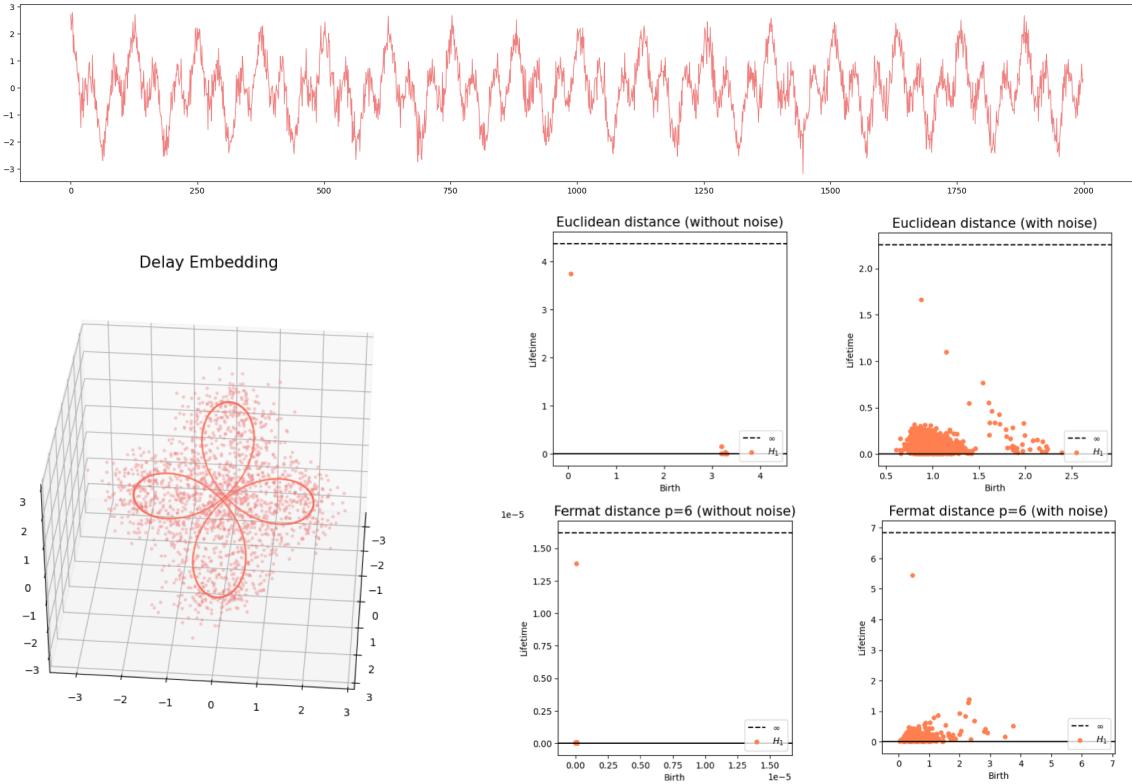


Figure 7: Top: Periodic signal with noise, defined as $f(t) = \cos(t) + \cos(3t)$ with additive Gaussian noise of variance 0.4. Bottom left: Delay embedding (projection 3d to the first coordinates) with the optimal values of the parameters, i.e $D = 8, \tau = 28$, according to the canonical heuristics (embedding of the signal without noise in dark orange). Bottom right: Persistence diagrams (degree 1 only) of the embedding of the signal without and with noise, computed using the Euclidean distance and Fermat distance for $p = 6$.

where $\mathcal{M}[0, t]$ is considered a metric subspace of $\mathcal{M} = \mathcal{M}[0, T]$ endowed with the population Fermat distance. Finally, compute

$$\frac{d_b(\text{dgm}(\text{Rips}(\mathbb{X}_i)), \text{dgm}(\text{Rips}(\mathbb{X}_{i-1})))}{t_i - t_{i-1}} \quad (8)$$

as an approximate the ‘first order derivative’ of (7). Shifts in patterns in the signal can be detected from the sample as peaks in the bottleneck distance between consecutive persistence diagrams.

Some applications of this technique follow below.

Example 7 (Anomaly detection in ECG) *The purpose of this example is to present a computational method of automated detection of abnormal heartbeats (arrhythmia) through the topological analysis of a delay embedding of ECG signals. We consider the record sel102*

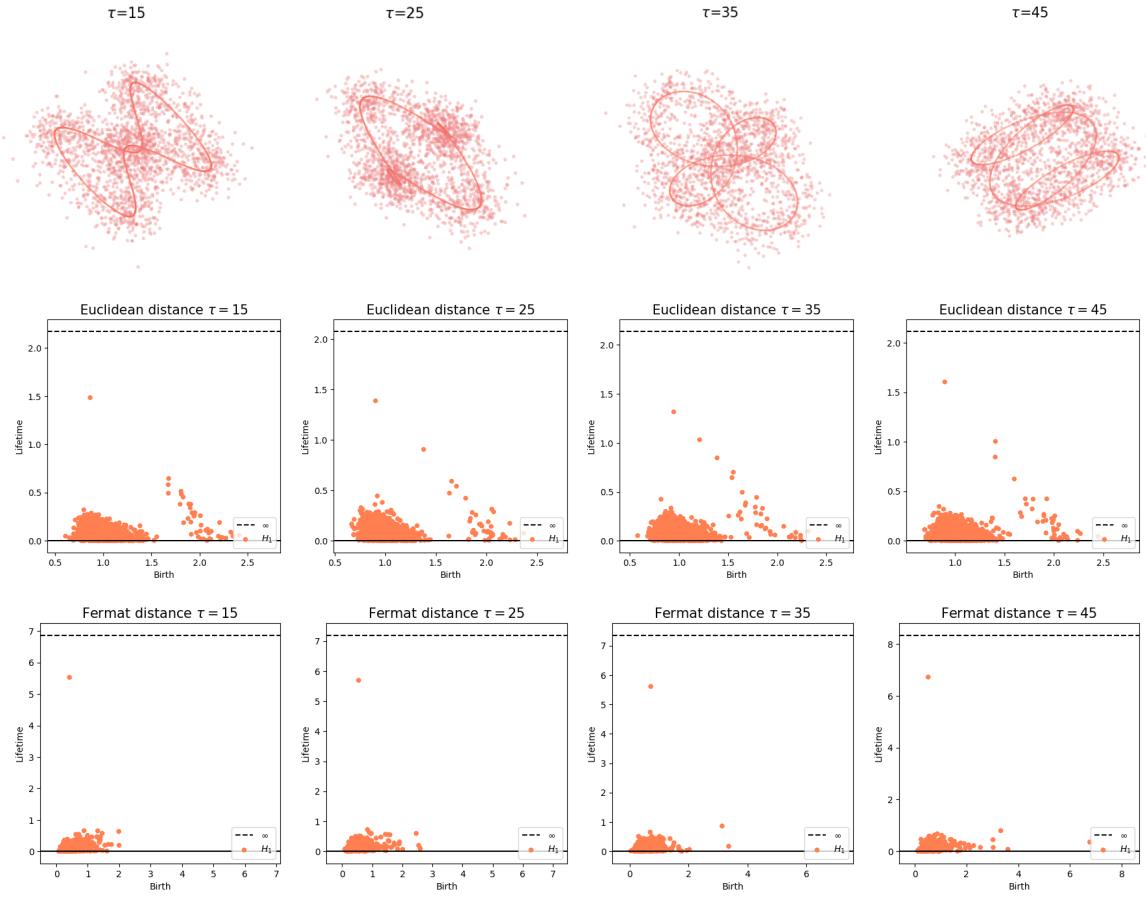


Figure 8: Top: Time-delay embeddings in \mathbb{R}^8 (projection 3d to the first coordinates) for $\tau = 15, 25, 35, 45$ of the signal $f(t) = \cos(t) + \cos(3t)$ with additive Gaussian noise of variance 0.4 (cf. Fig. 7). Bottom: Persistence diagrams (degree 1 only) using Euclidean distance and Fermat distance (for $p = 6$, but similar outputs are obtained for a range of values of p).

of the QT Database from the freely-available repository of medical research data PhysioNet MIT, Figure 9.

Regular heartbeats are characterized by a periodic pattern (Lilly and School, 2016, Ch.4). The delay embedding in \mathbb{R}^3 of a normal ECG has hence a cyclic topology induced by the periodic behavior of the time series (see Perea, 2019; Emrani et al., 2014). However, every time that an irregular heartbeat occurs, a new cycle arises in the embedding. We compute the associated persistence diagram for a normal period and for a period that includes an anomalous heartbeat. All delay embeddings were computed with a stride of $t = 2$, obtaining point clouds of up to ~ 3000 points from the original sample of size 6000. Persistent cycles in H_1 in diagrams computed using Euclidean distance are not in correspondence with the periodicity pattern and the anomaly. Indeed, at the periodic interval $[0, 4000]$ there are two salient generators for H_1 . On the contrary, by using Fermat distance, an initial cycle for

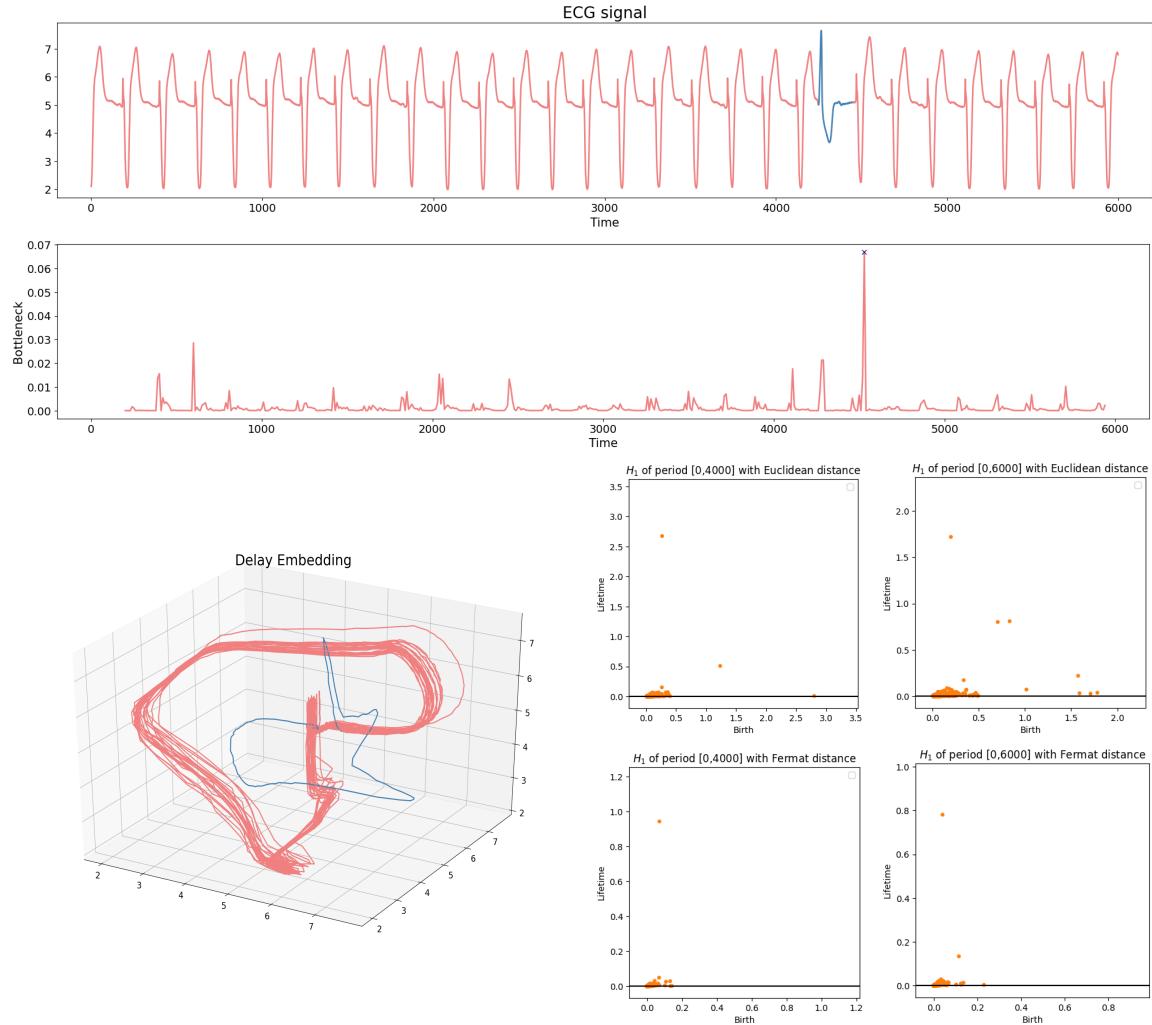


Figure 9: Top: ECG signal (anomaly in blue). Middle: Bottleneck distance between consecutive persistence diagrams associated to time evolving embeddings of the ECG signal. Bottom: Delay embedding in \mathbb{R}^3 with $\tau = 15$. The associated persistence diagrams at degree 1 using Euclidean distance and Fermat distance with $p = 2$ for the embedding of the signal in the periods of time $[0, 4000]$ and $[0, 6000]$.

the periodic pattern and a second cycle in the irregular period that accounts for the anomaly are distinctly detected (here, the choice of $p = 2$ is related to the weight we give to the density when computing Fermat distances; that is, we set p so that the exponent $\frac{p-1}{d}$ equals 1, where $d = 1$ is the dimension of the curve). Moreover, the moment immediately following the occurrence of the anomaly can be detected using persistent homology of time evolving delay embeddings. Indeed, the estimator (8) of the first derivative of the time evolving persistent diagrams features a prominent peak when the topology of the embedding changes. Lower peaks are also present as the result of the noisy real record.

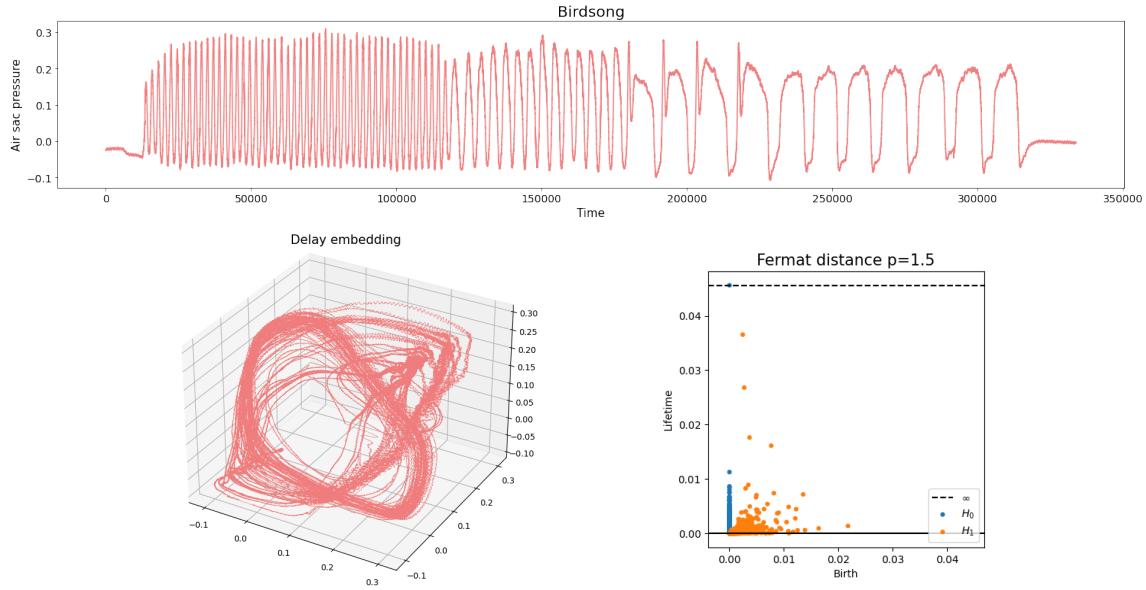


Figure 10: Top: Record of the air sac pressure of canary during a song. Bottom: Delay embedding in \mathbb{R}^3 with time delay $\tau = 500$ and its associated persistence diagram using Fermat distance with $p = 1.5$.

Example 8 (Pattern recognition in birdsongs) *During song production, canaries use a set of air sac pressure gestures with characteristic shapes to generate different patterns of sound (or syllables). Pressure patterns of different syllables constitute a diverse set: they can be either almost harmonic oscillations, high frequency fluctuations or oscillations presenting wiggles. The recognition of song syllables from the air sac pressure series is a well-studied problem in non-linear dynamical systems (Mindlin and Laje, 2006; Alonso et al., 2009).*

We provide a topological method to detect the number of different syllables in a canary song from the (noisy) record of the fluctuations of its air sac pressure $X(t)$, Figure 10 (data provided by the Laboratory of Dynamical Systems from the Department of Physics of the University of Buenos Aires). Given the time delay embedding of the time series $X(t)$ with $\tau = 500$ and $D = 3$, its associated persistence diagram computed using Fermat distance with $p = 1.5$ shows four prominent generators for the first homology group, which are in correspondence with the four different patterns observed in the time series (see Figure 11). Indeed, the embedding of each syllable is topologically a cycle (see Perea, 2019; Perea and Harer, 2015). However, this decomposition is not available beforehand so the study of the global topology of the embedding of the entire time series is necessary in order to analyze the complete song. Here, prior to the computation of the persistence diagram, we down-sampled the original time series at evenly spaced times with stride $t = 100$, obtaining a subsample of size ~ 3000 from the original $T \sim 300000$ points.

We can also detect the moments at which changes of syllables take place during the song. The estimator (8) of the first derivative of the path of persistence diagrams associated to the

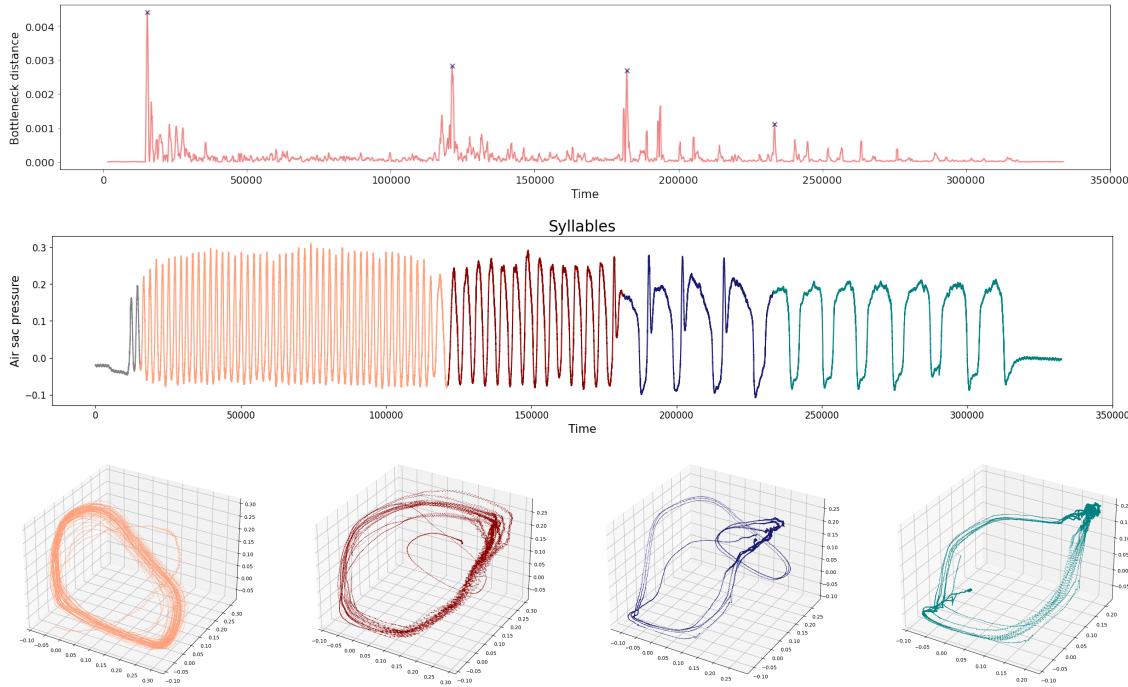


Figure 11: Top: Bottleneck distance between consecutive persistence diagrams associated to time evolving embeddings (moving average curve with window of time 500). Peaks are related to changes in the pattern of the air sac pressure record of the canary song. Bottom: Delay embedding of each detected syllable.

time evolving delay embeddings presents peaks followed by an exponential decay each time a new pattern arises, Figure 11.

5. Conclusions and Future Work

We introduced the use of density-based asymptotically intrinsic distances in point clouds to reconstruct the homology of a manifold from a noisy sample. In most of the standard approaches, persistent homology computed from Euclidean samples of manifolds lacks of two relevant properties: robustness to outliers and independence of the embedding in the ambient space. Whereas each of these properties has been studied separately in previous works, we present a simple method that is able to achieve both at the same time.

Our proposal is based on the use of Fermat distance when computing persistence diagrams of samples of manifolds. The key point is that, although this distance deforms the inherited geometry of the manifold, it produces intrinsic persistence diagrams that are more robust to outliers. Concretely, we provided rigorous proofs of convergence of the persistence diagrams of the associated metric spaces, robustness to a simple model of outliers and dependence of the persistence intervals on intrinsic (but not extrinsic) attributes of the underlying manifold. Furthermore, we showed experimentally that our technique is stable

under to a wider range of noisy situations, including real datasets. We intend to extend our results to more general models of outliers and noise in future works. Finally, a detailed comparison of our approach with other related methods, like DTM-filtrations and the use of Euclidean distance and the intrinsic k -NN distance in the construction of Vietoris-Rips filtrations, is also presented.

Acknowledgments

We are grateful to Luis Scoccola and Jeffrey Giansiracusa for many useful discussions and suggestions during the preparation of this article. We also acknowledge the anonymous reviewers and the associate editor for many helpful comments that greatly improved the manuscript. X. F. is a member of the Centre for Topological Data Analysis funded by the EPSRC grant EP/R018472/1. P. G. is partially supported by CONICET grant PIP 2021 11220200102825CO and UBACyT grant 20020190100293BA. G. M. is partially supported by PICT MAX PLANCK 4681 and PICT 00619.

Appendix A. Proof of Auxiliary Results

The purpose of this appendix is to present formal proofs of Proposition 5 and Lemma 8. Recall that $\mathcal{M} \subseteq \mathbb{R}^D$ is a closed submanifold of dimension $d \leq D$ and $\mathbb{X}_n \subseteq \mathcal{M}$ is an i.i.d. sample of size n with common density $f > 0$. Given $p > 1$, we set $\alpha = 1/(d + 2p)$.

Proposition 5 will be derived from Theorem 6 by Hwang et al. (2016). We start with a series of results to show that any segment that is part of any shortest path with respect to $d_{\mathbb{X}_n, p}$ is arbitrarily small with high probability for n large enough. This will allow us to prove that the sample Fermat distance uniformly well-approximates the power-weighted distance (1).

Proposition 18 *Given $b > 0$ and $\varepsilon > 0$, there exists $\theta > 0$ such that*

$$\mathbb{P} \left(\sup_{x,y} \left(\frac{n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y)}{d_{f,p}(x, y)} - \mu \right) > \varepsilon \right) \leq \exp(-\theta n^\alpha)$$

for n large enough, where the supremum is taken over all $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq b$.

Proof Given $\varepsilon > 0$ and $b > 0$, by Theorem 6 there exists $\theta > 0$ such that for every $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq b$,

$$\frac{n^{(p-1)/d} L_{\mathbb{X}_n, p}(x, y)}{d_{f,p}(x, y)} - \mu > \varepsilon \quad (9)$$

with probability at most $\exp(-\theta n^\alpha)$ (notice that here we set the sequence b_n to be constantly b).

Let $x, y \in \mathcal{M}$ and let $\gamma = (x_0, \dots, x_{k+1})$ be the shortest path between x, y with respect to $L_{\mathbb{X}_n, p}$. That is,

$$L_{\mathbb{X}_n, p}(x, y) = \sum_{i=0}^k d_{\mathcal{M}}(x_{i+1}, x_i)^p.$$

Since $|x_{i+1} - x_i| \leq d_{\mathcal{M}}(x_{i+1}, x_i)$,

$$L_{\mathbb{X}_n, p}(x, y) \geq \sum_{i=0}^k |x_{i+1} - x_i|^p \geq d_{\mathbb{X}_n, p}(x, y).$$

Thus, by (9), the inequality

$$\frac{n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y)}{d_{f, p}(x, y)} - \mu > \varepsilon$$

holds with probability bounded by $\exp(-\theta n^\alpha)$. ■

Corollary 19 Let $b_0 > 0$. Let $x, y \in \mathcal{M}$ be such that they belong to some minimal path between points in \mathcal{M} with respect to $d_{\mathbb{X}_n, p}$. Then,

$$\mathbb{P}(|x - y| > b_0) \leq \exp(-\theta n^\alpha)$$

for some constant $\theta > 0$, provided n is large enough.

Proof Fix $\varepsilon_0 > 0$. By Proposition 18, there exists a constant $\theta > 0$ such that

$$\mathbb{P}\left(\sup_{u, v} \frac{n^{(p-1)/d} d_{\mathbb{X}_n, p}(u, v)}{d_{f, p}(u, v)} > \mu + \varepsilon_0\right) \leq \exp(-\theta n^\alpha)$$

for all n sufficiently large, where the supremum is taken over $u, v \in \mathcal{M}$ such that $d_{\mathcal{M}}(u, v) \geq b_0$.

On the other hand, note that since \mathcal{M} is compact the diameter $\text{diam}_p(\mathcal{M})$ of \mathcal{M} with respect to the distance $d_{f, p}$ is finite. Hence,

$$\frac{d_{f, p}(u, v)}{n^{(p-1)/d}} (\mu + \varepsilon_0) \leq \frac{\text{diam}_p(\mathcal{M})}{n^{(p-1)/d}} (\mu + \varepsilon_0) \leq b_0^p$$

for all $u, v \in \mathcal{M}$ with $d_{\mathcal{M}}(u, v) \geq b_0$ and all n sufficiently large.

Suppose now that $x, y \in \mathcal{M}$ belong to some shortest path between points of \mathcal{M} with respect to $d_{\mathbb{X}_n, p}$, say u and v , but that $|x - y| > b_0$. Then, clearly $d_{\mathbb{X}_n, p}(u, v) \geq |x - y|^p$ and $d_{\mathcal{M}}(u, v) > b_0$ (since otherwise $d_{\mathbb{X}_n, p}(u, v) \leq |u - v|^p < b_0^p$). We remark here that x and y do not necessarily belong to the sample \mathbb{X}_n . From the previous computations, it follows that whenever n is large enough, with probability at least $1 - \exp(-\theta n^\alpha)$,

$$|x - y|^p \leq d_{\mathbb{X}_n, p}(u, v) \leq \frac{d_{f, p}(u, v)}{n^{(p-1)/d}} (\mu + \varepsilon_0) \leq b_0^p,$$

as we wanted to show. ■

Remark 20 (see Bernstein *et al.*, 2000, Corollary 4 or Boissonnat *et al.*, 2019, Lemma 3) Let (\mathcal{M}, g) be a smooth compact Riemannian manifold embedded in \mathbb{R}^D . Given $\delta > 0$, there exists $\varepsilon > 0$ such that for every $x, y \in \mathcal{M}$ with $|x - y| < \varepsilon$,

$$d_{\mathcal{M}}(x, y) \leq (1 + \delta)|x - y|.$$

We are now able to prove a new version of Theorem 6 in which the proposed estimator of $d_{f,p}$ is the sample Fermat distance (rather than the power-weighted shortest path).

Proposition 21 Fix $\varepsilon > 0$ and a sequence of positive real numbers $(b_n)_{n \geq 1}$ satisfying that $\frac{\log(n)}{nb_n^d} \rightarrow 0$ when $n \rightarrow \infty$. Then, for every $p > 1$, there exists $\theta > 0$ such that

$$\mathbb{P} \left(\sup_{x,y} \left| \frac{n^{(p-1)/d} d_{\mathbb{X}_n,p}(x, y)}{d_{f,p}(x, y)} - \mu \right| > \varepsilon \right) \leq \exp(-\theta(nb_n^d)^\alpha)$$

for n large enough, where the supremum is taken over $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq b_n$.

Proof Let $\delta > 0$ be a small number to be fixed later. The strategy of the proof consists of showing that, with probability exponentially high in $(nb_n^d)^\alpha$, $L_{\mathbb{X}_n,p}(x, y)$ and $d_{\mathbb{X}_n,p}(x, y)$ coincide up to a factor of $(1 + \delta)^p$ for all $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq b_n$. Once that is established, the proof follows readily by applying Theorem 6.

Notice in first place that by Remark 20, there exists $\eta > 0$ such that $d_{\mathcal{M}}(x, y) \leq (1 + \delta)|x - y|$ whenever $x, y \in \mathcal{M}$, $|x - y| < \eta$. By Corollary 19, we may assume that $|u - v| < \eta$ for every $u, v \in \mathcal{M}$ belonging to a minimal path with probability exponentially high in n^α . Let $x, y \in \mathcal{M}$ be two points with $d_{\mathcal{M}}(x, y) \geq b_n$. Since by our assumptions every segment in a shortest path from x to y with respect to $d_{\mathbb{X}_n,p}$ has Euclidean length at most η , it is not difficult to see that

$$d_{\mathbb{X}_n,p}(x, y) \leq L_{\mathbb{X}_n,p}(x, y) \leq (1 + \delta)^p d_{\mathbb{X}_n,p}(x, y). \quad (10)$$

Now, by Theorem 6, the probability that

$$\left| \frac{n^{(p-1)/d} L_{\mathbb{X}_n,p}(x, y)}{d_{f,p}(x, y)} - \mu \right| < \frac{\varepsilon}{2} \quad (11)$$

is exponentially high in $(nb_n^d)^\alpha$, provided n is large enough. We will check that for $\delta > 0$ sufficiently small, the desired inequality for $d_{\mathbb{X}_n,p}$ follows if we assume that the event from (11) occurs. It is clear by (10) and (11) that

$$\frac{n^{(p-1)/d} d_{\mathbb{X}_n,p}(x, y)}{d_{f,p}(x, y)} - \mu < \frac{\varepsilon}{2}.$$

As for the other inequality, notice that

$$-\frac{\varepsilon}{2} < (1 + \delta)^p \left(\frac{n^{(p-1)/d} d_{\mathbb{X}_n,p}(x, y)}{d_{f,p}(x, y)} - \mu \right) + ((1 + \delta)^p - 1)\mu.$$

Hence, for $\delta > 0$ small enough we have

$$-\varepsilon < \frac{n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y)}{d_{f,p}(x, y)} - \mu$$

as desired. \blacksquare

Finally, we promote the convergence of the sample Fermat distance from Proposition 21 to a *uniform* convergence in probability (that is, for any pair of points $x, y \in \mathcal{M}$ regardless of the distance between them). Such uniform convergence may be accomplished by choosing a sequence $(b_n)_{n \geq 1}$ which converges to 0 at an adequate rate. This step is instrumental in order to prove the Gromov–Hausdorff convergence of the sample metric spaces $(\mathbb{X}_n, d_{n,p})$ to $(\mathcal{M}, d_{f,p})$ (see Theorem 10 and its proof).

Proof [Proposition 5] Roughly, the strategy of the proof consists in bounding the quantity

$$|n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y) - \mu d_{f,p}(x, y)|$$

splitting in two cases according to whether the distance $d_{\mathcal{M}}(x, y)$ is greater than or smaller than some appropriately chosen sequence $b_n > 0$. More precisely, we will set $b_n = n^{-\lambda}$ for some $\lambda \in ((p-1)/pd, 1/d)$. Let $\varepsilon > 0$. Since $\lambda < 1/d$, clearly the sequence $\left(\frac{\log(n)}{nb_n^d}\right)_{n \geq 1}$ converges to 0 as n goes to infinity and hence, by Proposition 21 the bound

$$\left| \frac{n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y)}{d_{f,p}(x, y)} - \mu \right| > \varepsilon'$$

holds with probability at most $\exp(-\theta(nb_n^d)^\alpha) = \exp(-\theta n^{(1-\lambda d)\alpha})$ for some $\theta > 0$ and all $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq n^{-\lambda}$ provided n is large enough (here $\varepsilon' > 0$ is a small number to be determined). Denote by $\text{diam}(\mathcal{M})$ the diameter of \mathcal{M} with respect to the distance $d_{\mathcal{M}}$. Since $d_{f,p}(x, y) \leq m_f^{-(p-1)/d} d_{\mathcal{M}}(x, y) \leq m_f^{-(p-1)/d} \text{diam}(\mathcal{M})$, we see that the event

$$|n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y) - \mu d_{f,p}(x, y)| > m_f^{-(p-1)/d} \text{diam}(\mathcal{M}) \varepsilon'$$

also holds with probability bounded from above by $\exp(-\theta n^{(1-\lambda d)\alpha})$ for the same $\theta > 0$ as before, whenever $d_{\mathcal{M}}(x, y) \geq n^{-\lambda}$. By setting $\varepsilon' = \varepsilon(m_f^{-(p-1)/d} \text{diam}(\mathcal{M}))^{-1}$ we obtain the desired bound for $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \geq n^{-\lambda}$. For the remaining case, take $x, y \in \mathcal{M}$ satisfying $d_{\mathcal{M}}(x, y) \leq n^{-\lambda}$ and notice in first place that

$$d_{f,p}(x, y) \leq m_f^{-(p-1)/d} d_{\mathcal{M}}(x, y) \leq m_f^{-(p-1)/d} n^{-\lambda}.$$

Hence, for n sufficiently large, $\mu d_{f,p}(x, y) \leq \varepsilon/2$. On the other hand, since by definition of $d_{\mathbb{X}_n, p}$ it is

$$d_{\mathbb{X}_n, p}(x, y) \leq |x - y|^p \leq d_{\mathcal{M}}(x, y)^p \leq n^{-\lambda p},$$

we see that $n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y) \leq n^{(p-1)/d - \lambda p}$. The hypothesis on λ implies that the exponent of n in the last inequality is negative and thus $n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y) \leq \varepsilon/2$ provided n is

large. Summing up, we conclude that there exists n_0 such that for all $x, y \in \mathcal{M}$ with $d_{\mathcal{M}}(x, y) \leq n^{-\lambda}$ and $n \geq n_0$,

$$|n^{(p-1)/d} d_{\mathbb{X}_n, p}(x, y) - \mu d_{f, p}(x, y)| \leq \varepsilon,$$

which completes the proof of the proposition. \blacksquare

We turn now to the proof of Lemma 8, which follows ideas from Cuevas and Rodríguez-Casal (2004) and (Mémoli and Sapiro, 2005, Section 5).

Definition 22 (see Lee, 2018, Chapter 5) *The injectivity radius $\text{inj}(\mathcal{N})$ of a Riemannian manifold (\mathcal{N}, g) is defined as*

$$\text{inj}(\mathcal{N}) := \inf_{x \in \mathcal{N}} \text{inj}(\mathcal{N}, x),$$

where $\text{inj}(\mathcal{N}, x)$ is the largest radius for which the exponential map is a diffeomorphism.

Proof [Lemma 8] Since \mathcal{M} is compact, its injectivity radius $\text{inj}(\mathcal{M})$ is strictly positive. Then, by an inequality of Croke (see Croke, 1980, Proposition 14), there exists a constant $c = c(d) > 0$ such that every metric ball B in \mathcal{M} of radius $r < \frac{\text{inj}(\mathcal{M})}{2}$ has volume at least $c(d)r^d$. Since we can assume that $\kappa < 1$ without loss of generality, for all n sufficiently large we have $n^{(\kappa-1)/d} < \frac{\text{inj}(\mathcal{M})}{2}$. From this point, we follow the strategy from the proof of (Cuevas and Rodríguez-Casal, 2004, Theorem 3). Let P_n be the maximum number of disjoint balls of radius $\frac{n^{(\kappa-1)/d}}{4}$ contained in \mathcal{M} — this is known as *packing number*, see for example (Niyogi et al., 2008, Section 5) — and take $\{B_1, \dots, B_{P_n}\}$ a set of disjoint balls of radius $\frac{n^{(\kappa-1)/d}}{4}$ in \mathcal{M} . It is clear then that

$$P_n \leq \frac{\text{Vol}(\mathcal{M})}{\min_{1 \leq j \leq P_n} \text{Vol}(B_j)} \leq \frac{\text{Vol}(\mathcal{M})4^d}{c(d)} n^{1-\kappa},$$

for n so large that $n^{(\kappa-1)/d} < \frac{\text{inj}(\mathcal{M})}{2}$. Now, suppose that $x \in \mathcal{M}$ verifies $d_{\mathcal{M}}(x, \mathbb{X}_n) > n^{(\kappa-1)/d}$. Since the balls $2B_1, \dots, 2B_{P_n}$ cover \mathcal{M} (where $2B_j$ stands for the ball with the same center as B_j but with twice the radius) the distance from x to some center of these balls is at most $\frac{n^{(\kappa-1)/d}}{2}$ and thus there should be no point from the sample in some ball $2B_j$. A simple computation reveals that the probability that some random variable $\mathbf{x}_i \in \mathbb{X}_n$ does not belong to $2B_j$ is at most $1 - m_f \cdot \text{Vol}(2B_j)$. By the independence of the random variables $\mathbf{x}_1, \dots, \mathbf{x}_n$, if n is large enough

$$\mathbb{P}\left(\bigcap_{i=1}^n \{\mathbf{x}_i \notin 2B_j\}\right) \leq (1 - m_f \cdot \text{Vol}(2B_j))^n \leq (1 - m_f c(d)n^{\kappa-1})^n.$$

We conclude that

$$\mathbb{P}\left(\left\{\sup_{x \in \mathcal{M}} d_{\mathcal{M}}(x, \mathbb{X}_n) \geq n^{(\kappa-1)/d}\right\}\right) \leq \sum_{j=1}^{P_n} \mathbb{P}\left(\bigcap_{i=1}^n \{\mathbf{x}_i \notin 2B_j\}\right) \leq (1 - m_f c(d)n^{\kappa-1})^n P_n.$$

Since P_n grows at most like a polynomial in n , $(1 - m_f c(d)n^{\kappa-1})^n P_n \leq \exp(-\theta n^\kappa)$ for an appropriate $\theta > 0$ and n big enough, as we wanted to show. \blacksquare

References

- Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electron. J. Stat.*, 13(1):1359–1399, 2019.
- Michał Adamaszek and Henry Adams. The Vietoris-Rips complexes of a circle. *Pacific J. Math.*, 290(1):1–40, 2017.
- Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer, 2001.
- Leandro M. Alonso, Jorge A. Allende, Franz Goller, and Gabriel B. Mindlin. Low-dimensional dynamical model for the diversity of pressure patterns used in canary song. *Physical Review E*, 79(4):041929, 2009.
- Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. DTM-based filtrations. In *35th International Symposium on Computational Geometry*, volume 129 of *LIPICS. Leibniz Int. Proc. Inform.*, pages Art. No. 58, 15. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019.
- Aristas. Fermat package. <http://www.aristas.com.ar/fermat/>, 2018.
- Antonio Auffinger, Michael Damron, and Jack Hanson. *50 years of First-passage Percolation*, volume 68 of *Univ. Lect. Ser.* Providence, RI: American Mathematical Society (AMS), 2017. ISBN 978-1-4704-4183-8; 978-1-4704-4356-6.
- Mukund Balasubramanian and Eric L. Schwartz. The Isomap algorithm and topological stability. *Science*, 295 5552:7, 2002.
- Ulrich Bauer. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *J. Appl. Comput. Topol.*, 5(3):391–423, 2021. ISSN 2367-1726. doi: 10.1007/s41468-021-00071-5. URL <https://doi.org/10.1007/s41468-021-00071-5>.
- Paul Bendich, Taras Galkovskyi, and John Harer. Improving homology estimates with random walks. *Inverse Problems*, 27(12):124002, 14, 2011. ISSN 0266-5611.
- Mira Bernstein, Vin De Silva, John C. Langford, and Joshua B. Tenenbaum. Graph approximations to geodesics on embedded manifolds, 2000.
- Joan S. Birman and R. F. Williams. Knotted periodic orbits in dynamical systems. I. Lorenz’s equations. *Topology*, 22(1):47–82, 1983.
- Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and Topological Inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2018.
- Jean-Daniel Boissonnat, André Lieutier, and Mathijs Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *J. Appl. Comput. Topol.*, 3(1-2):29–58, 2019.

- Mickaël Buchet, Frédéric Chazal, Steve Y Oudot, and Donald R Sheehy. Efficient and robust persistent homology for measures. *Computational Geometry*, 58:70–96, 2016.
- Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.
- Frédéric Chazal and André Lieutier. Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees. *Comput. Geom.*, 40(2):156–170, 2008.
- Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, pages 237–246, 2009.
- Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11(6):733–751, 2011.
- Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geom. Dedicata*, 173:193–214, 2014.
- Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16: 3603–3635, 2015.
- Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The Structure and Stability of Persistence Modules*. SpringerBriefs in Mathematics. Springer, Cham, 2016.
- Jeff Cheeger and David G. Ebin. *Comparison Theorems in Riemannian Geometry*. North-Holland Publishing Co., Amsterdam-Oxford; American Elsevier Publishing Co., Inc., New York, 1975. North-Holland Mathematical Library, Vol. 9.
- Timothy Chu, Gary L. Miller, and Donald R. Sheehy. Exact computation of a manifold metric, via Lipschitz embeddings and shortest paths on a graph. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms*, pages 411–425. SIAM, Philadelphia, PA, 2020.
- Michael B. Cohen, Brittany Terese Fasy, Gary L. Miller, Amir Nayyeri, Donald R. Sheehy, and Ameya Velingker. Approximating nearest neighbor distances. In *Proceedings of the Algorithms and Data Structures Symposium*, 2015.
- David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007. ISSN 0179-5376.
- Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pages 1–6, 1987.
- Christopher B. Croke. Some isoperimetric inequalities and eigenvalue estimates. *Ann. Sci. École Norm. Sup. (4)*, 13(4):419–435, 1980.
- Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Adv. in Appl. Probab.*, 36(2):340–354, 2004.

- Michael Damron and Xuan Wang. Entropy reduction in Euclidean first-passage percolation. *Electron. J. Probab.*, 21:Paper No. 65, 23, 2016.
- Vin de Silva and R. Ghrist. Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7:339–358, 2007a.
- Vin de Silva and Robert Ghrist. Coverage in sensor networks via persistent homology. *Algebr. Geom. Topol.*, 7:339–358, 2007b.
- Herbert Edelsbrunner and John Harer. Persistent homology—a survey. In *Surveys on Discrete and Computational Geometry*, volume 453 of *Contemp. Math.*, pages 257–282. Amer. Math. Soc., Providence, RI, 2008.
- Herbert Edelsbrunner, David G. Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory*, 29(4):551–559, 1983.
- Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28(4):511–533, 2002. ISSN 0179-5376. Discrete and computational geometry and graph drawing (Columbia, SC, 2001).
- Saba Emrani, Thanos Gentimis, and Hamid Krim. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Processing Letters*, 21(4):459–463, 2014.
- Brittany T. Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014.
- Herbert Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93:418–491, 1959.
- Ximena Fernandez. Github repository: Intrinsic persistent homology. <https://github.com/ximenafernandez/intrinsicPH>, 2021a. Code for the computational examples.
- Ximena Fernandez. Intrinsic persistent homology. <https://www.youtube.com/watch?v=11P9ndiM60o>, 2021b. Prepared in the context of the *Tutorial-a-thon 2021*, organized by the Applied Algebraic Topology Research Network.
- Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33:1134–1140, Feb 1986.
- Rickard B. Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1553–1563. PMLR, 26–28 Aug 2020.
- Marian Gidea and Yuri Katz. Topological data analysis of financial time series: landscapes of crashes. *Phys. A*, 491:820–834, 2018.
- Robert Gilmore and Marc Lefranc. *The Topology of Chaos*. Wiley-Interscience [John Wiley & Sons], New York, 2002. Alice in Stretch and Squeezeland.

- Barbara Giunti, Guillaume Houry, and Michael Kerber. Average complexity of matrix reduction for clique filtrations. *arXiv preprint arXiv:2111.02125*, 2022.
- Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.
- Pablo Groisman, Matthieu Jonckheere, and Facundo Sapienza. Nonhomogeneous Euclidean first-passage percolation and distance learning. *Bernoulli*, 28(1):255–276, 2022.
- Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, 2002.
- Jean-Claude Hausmann. On the Vietoris-Rips complexes and a cohomology theory for metric spaces. In *Prospects in Topology (Princeton, NJ, 1994)*, volume 138 of *Ann. of Math. Stud.*, pages 175–188. Princeton Univ. Press, Princeton, NJ, 1995.
- C. Douglas Howard and Charles M. Newman. Euclidean models of first-passage percolation. *Probab. Theory Related Fields*, 108(2):153–170, 1997. ISSN 0178-8051.
- C. Douglas Howard and Charles M. Newman. Geodesics and spanning trees for Euclidean first-passage percolation. *Ann. Probab.*, 29(2):577–623, 2001.
- Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero, III. Shortest path through random points. *Ann. Appl. Probab.*, 26(5):2791–2823, 2016.
- Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45:3403–3411, Mar 1992.
- Jisu Kim, Jaehyeok Shin, Frédéric Chazal, Alessandro Rinaldo, and Larry Wasserman. Homotopy Reconstruction via the Čech Complex and the Vietoris-Rips Complex. In *SoCG 2020 - 36th International Symposium on Computational Geometry*, Zurich, Switzerland, June 2020.
- Janko Latschev. Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold. *Arch. Math. (Basel)*, 77(6):522–528, 2001.
- John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, 2018.
- Leonard S. Lilly and Harvard Medical School. *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty*. Wolters Kluwer, 2016.
- Anna Little, Daniel McKenzie, and James M. Murphy. Balancing geometry and density: path distances on high-dimensional data. *SIAM J. Math. Data Sci.*, 4(1):72–99, 2022.
- Edward N. Lorenz. Deterministic nonperiodic flow. *J. Atmospheric Sci.*, 20(2):130–141, 1963.
- Slobodan Maletić, Yi Zhao, and Milan Rajković. Persistent topological features of dynamical systems. *Chaos*, 26(5):053105, 14, 2016.

- Daniel Mckenzie and Steven Damelin. Power Weighted Shortest Paths for Clustering Euclidean Data. *Foundations of Data Science*, 1(3):307, 2019.
- Facundo Mémoli and Guillermo Sapiro. Distance functions and geodesics on submanifolds of \mathbb{R}^d and point clouds. *SIAM J. Appl. Math.*, 65(4):1227–1260, 2005. ISSN 0036-1399.
- Nikola Milosavljević, Dmitriy Morozov, and Primoz Skraba. Zigzag persistent homology in matrix multiplication time. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, SoCG ’11, page 216–225, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306829.
- Gabriel B. Mindlin and Rodrigo Laje. *The Physics of Birdsong*. Springer Science & Business Media, 2006.
- Laboratory for Computational Physiology MIT. Physionet databases. <https://physionet.org/about/database/>.
- Dmitriy Morozov. Persistence algorithm takes cubic time in worst case. *BioGeometry News, Dept. Comput. Sci., Duke Univ*, 2, 2005.
- Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.
- Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017.
- Jose A. Perea. Topological times series analysis. *Notices Amer. Math. Soc.*, 66(5):686–694, 2019.
- Jose A. Perea and John Harer. Sliding windows and persistence: an application of topological methods to signal analysis. *Found. Comput. Math.*, 15(3):799–838, 2015.
- Sajama and Alon Orlitsky. Estimating and computing density based distance metrics. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML ’05, page 760–767. Association for Computing Machinery, 2005.
- Facundo Sapienza, Pablo Groisman, and Matthieu Jonckheere. Weighted geodesic distance following Fermat’s principle. In *International Conference on Learning Representation*, 2018.
- Stephen Smale. Differentiable dynamical systems. *Bull. Amer. Math. Soc.*, 73:747–817, 1967. ISSN 0002-9904.
- Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980 (Coventry, 1979/1980)*, volume 898 of *Lecture Notes in Math.*, pages 366–381. Springer, Berlin-New York, 1981.
- Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2020.

- Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- Christopher J. Tralie and Jose A. Perea. (Quasi)periodicity quantification in video data, using topology. *SIAM J. Imaging Sci.*, 11(2):1049–1077, 2018.
- Pascal Vincent and Yoshua Bengio. Density-sensitive metrics and kernels. In *Snowbird Learning Workshop*, 2003.
- Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005. ISSN 0179-5376.