Statistical analysis and parameter selection for Mapper

Mathieu Carriere, Bertrand Michel, Steve Y. Oudot

Statistical analysis and parameter selection for Mapper

Mathieu Carrière, Bertrand Michel and Steve Oudot

May 30, 2017

Abstract

In this article, we study the question of the statistical convergence of the 1-dimensional Mapper to its continuous analogue, the Reeb graph. We show that the Mapper is an optimal estimator of the Reeb graph, which gives, as a byproduct, a method to automatically tune its parameters and compute confidence regions on its topological features, such as its loops and flares. This allows to circumvent the issue of testing a large grid of parameters and keeping the most stable ones in the brute-force setting, which is widely used in visualization, clustering and feature selection with the Mapper.

1 Introduction

In statistical learning, a large class of problems can be categorized into supervised or unsupervised problems. For supervised learning problems, an output quantity Y must be predicted or explained from the input measures X. On the contrary, for unsupervised problems there is no output quantity Y to predict and the aim is to explain and model the underlying structure or distribution in the data. In a sense, unsupervised learning can be thought of as extracting features from the data, assuming that the latter come with unstructured noise. Many methods in data sciences can be qualified as unsupervised methods, among the most popular examples are association methods, clustering methods, linear and non linear dimension reduction methods and matrix factorization to cite a few (see for instance Chapter 14 in Friedman et al. (2001)). Topological Data Analysis (TDA) has emerged in the recent years as a new field whose aim is to uncover, understand and exploit the topological and geometric structure underlying complex. and possibly high-dimensional data. Most of TDA methods can thus be qualified as unsupervised. In this paper, we study a recent TDA algorithm called Mapper which was first introduced in Singh et al. (2007).

Starting from a point cloud $\mathbb{X}_n$ sampled from a metric space $\mathcal{X}$, the idea of Mapper is to study the topology of the sublevel sets of a function $f : \mathbb{X}_n \to \mathbb{R}$ defined on the point cloud. The function f is called a filter function and it has to be chosen by the user. The construction of Mapper depends on the choice of a cover $\mathcal{I}$ of the image of f by open sets. Pulling back $\mathcal{I}$ through f gives an open cover of the domain $\mathbb{X}_n$. It is then refined into a connected cover by splitting each element into its various clusters using a clustering algorithm whose choice is left to the user. Then, the Mapper is defined as the nerve of the connected cover, having one vertex per element, one edge per pair of intersecting elements, and more generally, one k-simplex per non-empty $(k+1)$-fold intersection.

In practice, the Mapper has two major applications. The first one is data visualization and clustering. Indeed, when the cover $\mathcal{I}$ is minimal, the Mapper provides a visualization of the data in the form of a graph whose topology reflects that of the data. As such, it brings additional

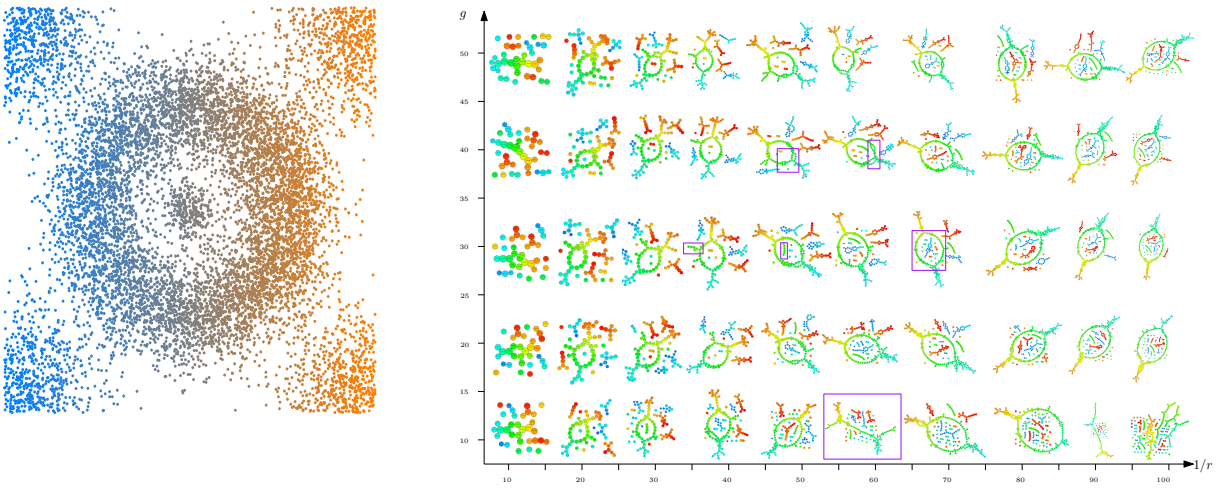

Figure 1: Bunch of Mappers computed with various parameters. Left: crater dataset. Right: outputs of Mapper with various parameters. One can see that for some Mappers, (the ones with purple squares), topological features suddenly appear and disappear. These are discretization artifacts, that we overcome in this article by appropriately tuning the parameters.

information to the usual clustering algorithms by identifying *flares* and *loops* that outline potentially remarkable subpopulations in the various clusters. See e.g. Yao et al. (2009); Lum et al. (2013); Sarikonda et al. (2014); Hinks et al. (2015) for examples of applications. The second application of Mapper deals with feature selection. Indeed, each feature of the data can be evaluated on its ability to discriminate the interesting subpopulations mentioned above (flares, loops) from the rest of the data, using for instance Kolmogorov-Smirnov tests. See e.g. Lum et al. (2013); Nielson et al. (2015); Rucco et al. (2015) for examples of applications.

Unsupervised methods generally depend on parameters that need to be chosen by the user. For instance, the number of selected dimensions for dimension reduction methods or the number of clusters for clustering methods have to be chosen. Contrarily to supervised problems, it is tricky to evaluate the output of unsupervised methods and thus to select parameters. This situation is highly problematic with Mapper since, as for many TDA methods, it is not robust to outliers. This major drawback of Mapper is an important obstacle to its use in Exploratory Data Analysis with non trivial datasets. This phenomenon is illustrated for instance in Figure 1 on a dataset that we study further in Section 5. The only answer proposed to this drawback in the literature consists in selecting parameters in a range of values for which the Mapper seems to be stable—see for instance Nielson et al. (2015). We believe that such an approach is not satisfactory because it does not provide statistical guarantees on the inferred Mapper.

Our main goal in this article is to provide a statistical method to tune the parameters of Mapper automatically. To select parameters for Mapper, or more generally to evaluate the significance of topological features provided by Mapper, we develop a rigorous statistical framework for the convergence of the Mapper. This contribution is made possible by the recent work (in a deterministic setting) of Carrière and Oudot (2016) about the structure and the stability of the Mapper. In this article, the authors explicit a way to go from the input space to the Mapper using small perturbations. We build on this relation between the input space and its Mapper to show that the

Mapper is itself a measurable construction. In Carrière and Oudot (2016), the authors also show that the topological structure of the Mapper can actually be predicted from the cover $\mathcal{I}$ by looking at appropriate *signatures* that take the form of *extended persistence diagrams*. In this article, we use this observation, together with an approximation inequality, to show that the Mapper, computed with a specific set of parameters, is actually an optimal estimator of its continuous analogue, the so-called *Reeb graph*. Moreover, these specific parameters act as natural candidates to obtain a reliable Mapper with no artifacts, avoiding the computational cost of testing millions of candidates and selecting the most stable ones in the brute-force setting of many practitioners. Finally, we also provide methods to assess the stability and compute confidence regions for the topological features of the Mapper. We believe that this set of methods open the way to an accessible and intuitive utilization of Mapper for non expert researchers in applied topology.

Section 2 presents the necessary background on the Reeb graph and Mapper, and it also gives an approximation inequality—Theorem 2.7—for the Reeb graph with the Mapper. From this approximation result, we derive rates of convergences as well as candidate parameters in Section 3, and we show how to get confidence regions in Section 4. Section 5 illustrates the validity of our parameter tuning and confidence regions with numerical experiments on smooth and noisy data.

2 Approximation of a Reeb graph with Mapper

2.1 Background on the Reeb graph and Mapper

We start with some background on the Reeb graph and Mapper. In particular, we present the specific Mapper algorithm that we study in this article.

Reeb graph. Let $\mathcal{X}$ be a topological space and let $f : \mathcal{X} \to \mathbb{R}$ be a continuous function. Such a function on $\mathcal{X}$ is called a *filter function* in the following. Then, we define the equivalence relation $\sim_f$ as follows: for all x and x' in $\mathcal{X}$, x and x' are in the same class ($x \sim_f x'$) if and only if x and x' belong to the same connected component of $f^{-1}(y)$, for some y in the image of f.

Definition 2.1. *The* Reeb graph $\mathrm{R}_f(\mathcal{X})$ *of* $\mathcal{X}$ *computed with the filter function* f *is the quotient space* $\mathcal{X}/\sim_f$ *endowed with the quotient topology.*

See Figure 2 for an illustration. Note that, since f is constant on equivalence classes, there is an induced map $f_{\mathrm{R}} : \mathrm{R}_f(\mathcal{X}) \to \mathbb{R}$ such that $f = f_{\mathrm{R}} \circ \pi$, where π is the quotient map $\mathcal{X} \to \mathrm{R}_f(\mathcal{X})$.

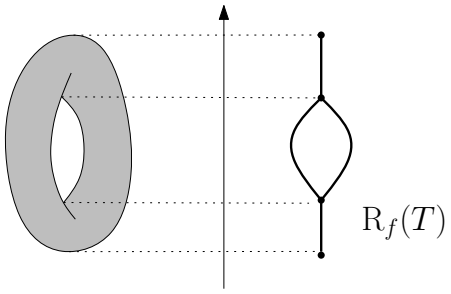

Figure 2: Example of Reeb graph computed on the torus T with the height function f.

The topological structure of a Reeb graph can be described if the pair $(\mathcal{X}, f)$ is regular enough. From now on, we will assume that the filter function $f : \mathcal{X} \to \mathbb{R}$ is *Morse-type*. Morse-type functions are generalizations of classical Morse functions that share some of their properties without having to be differentiable (nor even defined over a smooth manifold):

Definition 2.2. *A continuous real-valued function f on a compact space $\mathcal{X}$ is* of Morse type *if:*

(i) *There is a finite set* $\mathrm{Crit}(f) = \{a_1 < ... < a_n\}$*, called the set of* critical values*, such that over every open interval $(a_0 = -\infty, a_1), ..., (a_i, a_{i+1}), ..., (a_n, a_{n+1} = +\infty)$ there is a compact and locally connected space $\mathcal{Y}_i$ and a homeomorphism $\mu_i : \mathcal{Y}_i \times (a_i, a_{i+1}) \to \mathcal{X}^{(a_i, a_{i+1})}$ s.t. $\forall i = 0, ..., n, f|_{\mathcal{X}^{(a_i, a_{i+1})}} = \pi_2 \circ \mu_i^{-1}$, where π_2 is the projection onto the second factor;*

(ii) *$\forall i = 1, ..., n-1, \mu_i$ extends to a continuous function $\bar{\mu}_i : \mathcal{Y}_i \times [a_i, a_{i+1}] \to \mathcal{X}^{[a_i, a_{i+1}]}$ – similarly μ_0 extends to $\bar{\mu}_0 : \mathcal{Y}_0 \times (-\infty, a_1] \to \mathcal{X}^{(-\infty, a_1]}$ and μ_n extends to $\bar{\mu}_n : \mathcal{Y}_n \times [a_n, +\infty) \to \mathcal{X}^{[a_n, +\infty)}$;*

(iii) *Each levelset $\mathcal{X}^t$ has a finitely-generated homology.*

Key fact 1a: For $f : \mathcal{X} \to \mathbb{R}$ a Morse-type function, the Reeb graph $\mathrm{R}_f(\mathcal{X})$ is a multigraph.

For our purposes, in the following we further assume that $\mathcal{X}$ is a smooth and compact submanifold of $\mathbb{R}^D$. The space of Reeb graphs computed with Morse-type functions over such spaces is denoted $\mathcal{R}$ in this article.

Mapper. The Mapper is introduced in Singh et al. (2007) as a statistical version of the Reeb graph $\mathrm{R}_f(\mathcal{X})$ in the sense that it is a discrete and computable approximation of the Reeb graph computed with some filter function. Assume that we observe a point cloud $\mathbb{X}_n = \{X_1, \ldots, X_n\} \subset \mathcal{X}$ with known pairwise dissimilarities. A filter function is chosen and can be computed on each point of $\mathbb{X}_n$. The generic version of the Mapper algorithm on $\mathbb{X}_n$ computed with the filter function f can be summarized as follows:

1. Cover the range of values $\mathbb{Y}_n = f(\mathbb{X}_n)$ with a set of consecutive intervals $I_1, \ldots, I_S$ which overlap.

2. Apply a clustering algorithm to each pre-image $f^{-1}(I_s)$, $s \in \{1, ..., S\}$. This defines a *pullback cover* $\mathcal{C} = \{\mathcal{C}_{1,1}, \ldots, \mathcal{C}_{1,k_1}, \ldots, \mathcal{C}_{S,1}, \ldots, \mathcal{C}_{S,k_S}\}$ of the point cloud $\mathbb{X}_n$.

3. The Mapper is then the *nerve* of $\mathcal{C}$. Each vertex $v_{s,k}$ of the Mapper corresponds to one element $\mathcal{C}_{s,k}$ and two vertices $v_{s,k}$ and $v_{s',k'}$ are connected if and only if $\mathcal{C}_{s,k} \cap \mathcal{C}_{s',k'}$ is not empty.

Even for one given filter function, many versions of the Mapper algorithm can be proposed depending on how one chooses the intervals that cover the image of f, and which method is used to cluster the pre-images. Moreover, note that the Mapper can be defined as well for continuous spaces. The definition is strictly the same except for the clustering step, since the connected components of each pre-image $f^{-1}(I_s)$, $s \in \{1, ..., S\}$ are now well-defined. See Figure 3.

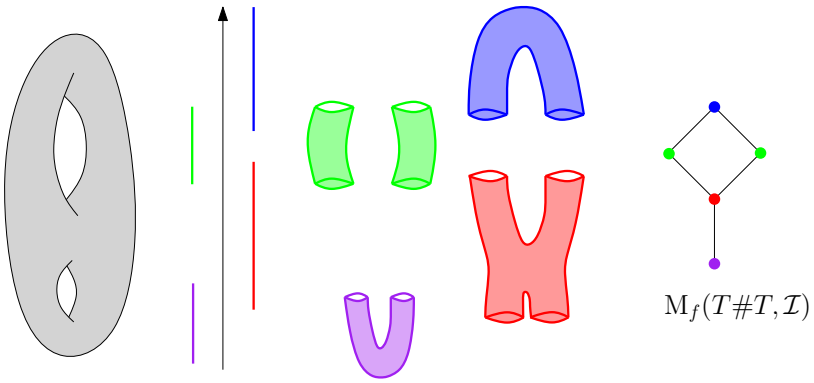

Figure 3: Example of Mapper computed on the double torus $T\#T$ with the height function f and a cover $\mathcal{I}$ of its range with four open intervals.

Our version of Mapper. In this article, we focus on a Mapper algorithm that uses neighborhood graphs. Of course, more sophisticated versions of Mapper can be used in practice but then the statistical analysis is more tricky. We assume that there exists a distance on $\mathbb{X}_n$ and that the matrix of pairwise distances is available. First, from the distance matrix we compute the 1-skeleton of a Rips complex with parameter δ, i.e. the δ-neighborhood graph built on top of $\mathbb{X}_n$. This objet plays the role of an approximation of the underlying and unknown metric space $\mathcal{X}$ on which the data are sampled. Second, given $\mathbb{Y}_n = f(\mathbb{X}_n)$ the set of filter values, we choose a regular cover of $\mathbb{Y}_n$ with open intervals, where no more than two intervals can intersect at a time. More precisely, we use open intervals with same length r: $\forall s \in \{1, \dots, S\}$,

$$r = \ell(I_s)$$

where ℓ is the Lebesgue measure on $\mathbb{R}$. The overlap g between two consecutive intervals is also a fixed constant: $\forall s \in \{1, \dots, S-1\}$,

$$0 < g = \frac{\ell(I_s \cap I_{s+1})}{\ell(I_s)} < \frac{1}{2}.$$

The parameters g and r are generally called the *gain* and the *resolution* in the literature on the Mapper algorithm. Finally, for the clustering step, we simply consider the connected components of the pre-images $f^{-1}(I_s)$ that are induced by the 1-skeleton of the Rips complex. The corresponding Mapper is denoted $\mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$ or M_n for short in the following. When dealing with a continuous space $\mathcal{X}$, there is no need to compute a neighborhood graph since the connected components are well-defined, so we let $\mathrm{M}_{r,g}(\mathcal{X}, f)$ denote such a Mapper.

Key fact 1b: The Mapper $\mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$ is a combinatorial graph.

Moreover, following Carrière and Oudot (2015), we can define a function on the nodes of M_n as follows.

Definition 2.3. *Let v be a node of M_n, i.e. v represents a connected component of $f^{-1}(I_s)$ for some $s \in \{1, \dots, S\}$. Then, we let*

$$f_{\mathcal{I}}(v) := \mathrm{mid}(\tilde{I}_s),$$

where $\tilde{I}_s := I_s \setminus (I_s \cap I_{s-1}) \cup (I_s \cap I_{s+1})$ and $\mathrm{mid}(\tilde{I}_s)$ denotes the midpoint of the interval $\tilde{I}_s$.

Filter functions. In practice, it is common to choose filter functions that are coordinate-independent, in order to avoid depending on solid transformations of the data like rotations or translations. The two most common filters that are used in the literature are:

- the *eccentricity*: $x \mapsto \sup_{y \in \mathcal{X}} d(x, y)$,
- the eigenfunctions given by a Principal Component Analysis of the data.

2.2 Extended persistence signatures and the persistence metric d_Δ

In this section, we introduce *extended persistence* and its associated metric, the *bottleneck distance*, which we will use later to compare Reeb graphs and Mappers.

Extended persistence. Given any graph $G = (V, E)$ and a function attached to its nodes $f : V \to \mathbb{R}$, the so-called *extended persistence diagram* $\mathrm{Dg}(G, f)$ is a multiset of points in the Euclidean plane $\mathbb{R}^2$ that can be computed with *extended persistence theory*. Each of the diagram points has a specific *type*, which is either Ord_0, Rel_1, Ext_0^+ or Ext_1^-. We refer the reader to Appendix C for formal definitions and further details about extended persistence. A rigorous connexion between the Mapper and the Reeb graph was drawn recently by Carrière and Oudot (2016), who show how extended persistence provides a relevant and efficient framework to compare a Reeb graph with a Mapper. We summarize below the main points of this work in the perspective of the present article.

Topological dictionary. Given a topological space $\mathcal{X}$ and a Morse-type function $f : \mathcal{X} \to \mathbb{R}$, there is a nice interpretation of $\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_\mathrm{R})$ in terms of the structure of $\mathrm{R}_f(\mathcal{X})$. Orienting the Reeb graph vertically so f_R is the height function, we can see each connected component of the graph as a trunk with multiple branches (some oriented upwards, others oriented downwards) and holes. Then, one has the following correspondences, where the *vertical span* of a feature is the span of its image by f_R:

- The vertical spans of the trunks are given by the points in $\mathrm{Ext}_0^+(\mathrm{R}_f(\mathcal{X}), f_\mathrm{R})$;
- The vertical spans of the branches that are oriented downwards are given by the points in $\mathrm{Ord}_0(\mathrm{R}_f(\mathcal{X}), f_\mathrm{R})$;
- The vertical spans of the branches that are oriented upwards are given by the points in $\mathrm{Rel}_1(\mathrm{R}_f(\mathcal{X}), f_\mathrm{R})$;
- The vertical spans of the holes are given by the points in $\mathrm{Ext}_1^-(\mathrm{R}_f(\mathcal{X}), f_\mathrm{R})$.

These correspondences provide a dictionary to read off the structure of the Reeb graph from the corresponding extended persistence diagram. See Figure 4 for an illustration.

Note that it is a bag-of-features type descriptor, taking an inventory of all the features (trunks, branches, holes) together with their vertical spans, but leaving aside the actual layout of the features. As a consequence, it is an incomplete descriptor: two Reeb graphs with the same persistence diagram may not be isomorphic.

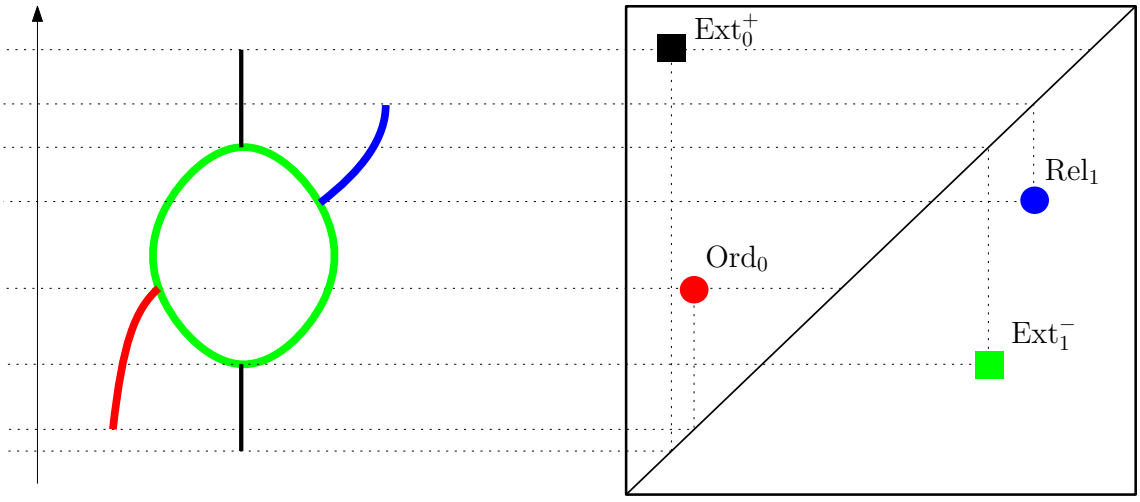

Figure 4: Example of correspondences between topological features of a graph and points in its corresponding extended persistence diagram. Note that ordinary persistence is unable to detect the blue upwards branch.

Bottleneck distance. We now define the commonly used metric between persistence diagrams.

Definition 2.4. *Given two persistence diagrams D, D', a* partial matching *between D and D' is a subset Γ of $D \times D'$ such that:*

$$\forall p \in D, \text{ there is at most one } p' \in D' \text{ s.t. } (p, p') \in \Gamma,$$

$$\forall p' \in D', \text{ there is at most one } p \in D \text{ s.t. } (p, p') \in \Gamma.$$

Furthermore, Γ must match points of the same type (ordinary, relative, extended) and of the same homological dimension only. Let Δ be the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$. The cost *of Γ is:*

$$cost(\Gamma) = \max\left\{\max_{p \in D} \delta_D(p), \max_{p' \in D'} \delta_{D'}(p')\right\},$$

where

$$\delta_D(p) = \|p - p'\|_\infty \text{ if } \exists p' \in D' \text{ s.t. } (p, p') \in \Gamma, \text{ otherwise } \delta_D(p) = \inf_{q \in \Delta} \|p - q\|_\infty,$$

$$\delta_{D'}(p') = \|p - p'\|_\infty \text{ if } \exists p \in D \text{ s.t. } (p, p') \in \Gamma, \text{ otherwise } \delta_{D'}(p') = \inf_{q \in \Delta} \|p' - q\|_\infty.$$

Definition 2.5. *Let D, D' be two persistence diagrams. The* bottleneck distance *between D and D' is:*

$$d_\Delta(D, D') = \inf_\Gamma \text{cost}(\Gamma),$$

where Γ ranges over all partial matchings between D and D'.

Note that d_Δ is only a pseudometric, not a true metric, because points lying on Δ can be left unmatched at no cost.

Definition 2.6. *Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two combinatorial graphs with real-valued functions $f_1 : V_1 \to \mathbb{R}$ and $f_2 : V_2 \to \mathbb{R}$ attached to their nodes. The* persistence metric d_Δ *between the pairs (G_1, f_1) and (G_2, f_2) is:*

$$d_\Delta(G_1, G_2) := d_\Delta\left(\mathrm{Dg}(G_1, f_1), \mathrm{Dg}(G_2, f_2)\right).$$

For a Morse-type function f defined on $\mathcal{X}$ and for a finite point cloud $\mathbb{X}_n \subset \mathcal{X}$, we can thus consider $\mathrm{Dg}(\mathrm{R}_f(\mathcal{X})) := \mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}})$ and $\mathrm{Dg}(\mathrm{M}_n) := \mathrm{Dg}(\mathrm{M}_n, f_{\mathcal{I}})$, with $f_{\mathcal{I}}$ as in Definition 2.3. In this context the bottleneck distance $d_\Delta(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n) = d_\Delta(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X})), \mathrm{Dg}(\mathrm{M}_n))$ is well defined and we use this quantity to assess if the Mapper M_n is a good approximation of the Reeb graph $\mathrm{R}_f(\mathcal{X})$. Moreover, note that, even though d_Δ is only a pseudometric, it has been shown to a be a true metric *locally* for Reeb graphs by Carrière and Oudot (2017).

As noted in Carrière and Oudot (2015), the choice of $f_{\mathcal{I}}$ is in some sense arbitrary since any function defined on the nodes of the Mapper that respects the ordering of the intervals of $\mathcal{I}$ carries the same information in its extended persistence diagram. To avoid this issue, Carrière and Oudot (2016) define a pruned version of $\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}})$ as a canonical descriptor for the Mapper. The problem is that computing this canonical descriptor requires to know the critical values of f_{R} beforehand. Here, by considering $\mathrm{Dg}(\mathrm{M}_n, f_{\mathcal{I}})$ instead, the descriptor becomes computable. Moreover, one can see from the proofs in the Appendix that the canonical descriptor and its arbitrary version actually enjoy the same rate of convergence, up to some constant.

Geometric quantity. Let $\pi_\Delta : \mathbb{R}^2 \to \mathbb{R}^2$ be the projection onto Δ. We define $e_{\min}(\mathcal{X}, f)$ as the smallest distance to the diagonal (in the ℓ_∞ norm) of the points of $\mathrm{Ext}_1^-(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}})$:

$$e_{\min} = \min\left\{\|p - \pi_\Delta(p)\|_\infty \mid p \in \mathrm{Ext}_1^-(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}})\right\}.$$

Intuitively, $e_{\min}$ is the size of the smallest loop one can find in $\mathcal{X}$. Hence, the larger $e_{\min}(\mathcal{X}, f)$, the smoother $\mathcal{X}$. This quantity plays an important role in the assumptions of our approximation result—see Theorem 2.7 in Section 2.3.

2.3 An approximation inequality for Mapper

We are now ready to give the key ingredient of this paper to derive a statistical analysis of the Mapper in Euclidean spaces. The ingredient is an upper bound on the bottleneck distance between the Reeb graph of a pair $(\mathcal{X}, f)$ and the Mapper computed with the same filter function f and a specific cover $\mathcal{I}$ of a sampled point cloud $\mathbb{X}_n \subset \mathcal{X}$. From now on, it is assumed that the underlying space $\mathcal{X}$ is a compact submanifold of dimension d embedded in $\mathbb{R}^D$, and that the filter function f is Morse-type on $\mathcal{X}$.

Regularity of the filter function. Intuitively, approximating a Reeb graph computed with a filter function f that has large variations is more difficult than for a smooth filter function, for some notion of regularity that we now specify. Our result is given in a general setting by considering the modulus of continuity of f. In our framework, f is assumed to be Morse-type and thus uniformly continuous on the compact set $\mathcal{X}$. Following for instance Section 6 in DeVore and Lorentz (1993), we define the (exact) modulus of continuity of f by

$$\omega_f(\delta) := \sup_{\|x - x'\| \leq \delta} |f(x) - f(x')|$$

for any $\delta > 0$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^D$. Then ω_f satisfies :

1. $\omega_f(\delta) \to \omega(0) = 0$ as $\delta \to 0$;
2. ω_f is non negative and non-decreasing on $\mathbb{R}^+$;
3. ω_f is subadditive : $\omega_f(\delta_1 + \delta_2) \leq \omega_f(\delta_1) + \omega_f(\delta_2)$ for any δ_1, $\delta_2 > 0$;
4. ω_f is continous on $\mathbb{R}^+$.

We say that a function ω defined on $\mathbb{R}^+$ is a modulus of continuity if it satisfies the four properties above and we say that it is a modulus of continuity of f if, in addition, we have

$$|f(x) - f(x')| \leq \omega(\|x - x'\|),$$

for any $x, x' \in \mathcal{X}$.

Theorem 2.7. *Assume that $\mathcal{X}$ is a compact submanifold of dimension d in $\mathbb{R}^D$ with positive reach rch and positive convexity radius ρ. Let $\mathbb{X}_n$ be a point cloud of n points, all lying in $\mathcal{X}$. Assume that the filter function f is Morse-type on $\mathcal{X}$ with $e_{\min} = e_{\min}(\mathcal{X}, f) > 0$ and $p_{\min} = p_{\min}(\mathcal{X}, f) > 0$. Let ω be a modulus of continuity of f. If the three following conditions hold:*

$$\delta \leq \frac{1}{4}\min\{rch, \rho\} \text{ and } \omega(\delta) \leq \frac{1}{2}e_{\min}, \tag{1}$$

$$\max\{|f(X) - f(X')| : X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta\} \leq gr, \tag{2}$$

$$4d_{\mathrm{H}}(\mathcal{X}, \mathbb{X}_n) \leq \delta, \tag{3}$$

then the Mapper $\mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$ with parameters r, g and δ is such that:

$$d_\Delta\left(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n\right) \leq \frac{r}{2} + 2\omega(\delta). \tag{4}$$

Remark 2.8. *Studying the MultiNerve Mapper—as defined in Carrière and Oudot (2015)—instead of the Mapper allows to weaken Assumption* (2) *since gr can be replaced by r in the corresponding equation.*

Analysis of the hypotheses. On the one hand, the scale parameter of the Rips complex could not be smaller than the approximation error corresponding to the Hausdorff distance between the sample and the underlying space $\mathcal{X}$ (Assumption (3)). On the other hand, it must be smaller than the reach and convexity radius to provide a correct estimation of the geometry and topology of $\mathcal{X}$ (Assumption (1)). The quantity gr corresponds to the minimum scale at which the filter's codomain is analyzed. This minimum resolution has to be compared with the regularity of the filter at scale δ (Assumption (2)). Indeed the pre-images of a filter with strong variations will be more difficult to analyze than when the filter does not vary too fast.

Analysis of the upper bound. The upper bound given in (4) makes sense in that the approximation error is controlled by the resolution level in the codomain and by the regularity of the filter. If one uses a filter with strong variations, or if the grid in the codomain has a too rough resolution, then the approximation will be poor. On the other hand, a sufficiently dense sampling is required in order to take r small, as prescribed in the assumptions.

Lipschitz filters. A large class of filters used for Mapper are actually Lipschitz functions and of course, in this case, one can take $\omega(\delta) = c\delta$ for some positive constant c. In particular, $c = 1$ for linear projections (PCA, SVD, Laplacian or coordinate filter for instance). The distance to a measure (DTM) is also a 1-Lipschitz function, see Chazal et al. (2011). On the other hand, the modulus of continuity of filter functions defined from estimators, e.g. density estimators, is less obvious although still well-defined.

Filter approximation. In some situations, the filter function $\hat{f}$ used to compute the Mapper is only an approximation of the filter function f with which the Reeb graph is computed. In this context, the pair $(\mathbb{X}_n, \hat{f})$ appears as an approximation of the pair $(\mathcal{X}, f)$. The following result is directly derived from Theorem 2.7 and Theorem 6.1 in Carrière and Oudot (2016) (that derives stability for Mappers building on the stability theorem of extended persistence diagrams proved by Cohen-Steiner et al. (2009)):

Corollary 2.9. *Let $\hat{f} : \mathcal{X} \to \mathbb{R}$ be a Morse-type filter function approximating f. Assume that Assumptions (1) and (3) of Theorem 2.7 are satisfied, and assume moreover that*

$$\max\{\{|f(X) - f(X')|, |\hat{f}(X) - \hat{f}(X')|\} \ : \ X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta\} \leq gr. \tag{5}$$

Then, the Mapper $\hat{\mathrm{M}}_n := \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \hat{f}(\mathbb{X}_n))$ built on $\mathbb{X}_n$ with filter function $\hat{f}$ and parameters r, g, δ satisfies:

$$d_\Delta(\mathrm{R}_f(\mathcal{X}), \hat{\mathrm{M}}_n) \leq \frac{3r}{2} + 2\omega(\delta) + \max_{1 \leq i \leq n} |f(X_i) - \hat{f}(X_i)|.$$

3 Statistical Analysis of Mapper

From now on, the set of observations $\mathbb{X}_n$ is assumed to be composed of n independent points $X_1, ..., X_n$ sampled from a probability distribution $\mathbb{P}$ in $\mathbb{R}^D$ (endowed with its Borel algebra). We assume that each point X_i comes with a filter value which is represented by a random variable Y_i. Contrarily to the X_i's, the filter values Y_i's are not necessarily independent. In the following, we consider two different settings: in the first one, $Y_i = f(X_i)$, where the filter f is a deterministic function, in the second one, $Y_i = \hat{f}(X_i)$ where $\hat{f}$ is an estimator of the filter function f. In the latter case, the Y_i's are obviously dependent. We first provide the following Proposition, whose proof is deferred to Appendix A.4, which states that computing probabilities on the Mapper makes sense:

Proposition 3.1. *For any fixed choice of parameters r, g, δ and for any fixed $n \in \mathbb{N}$, the function*

$$\Phi : \left\{ \begin{array}{ccc} (\mathbb{R}^D)^n \times \mathbb{R}^n & \to & \mathcal{R} \\ (\mathbb{X}_n, \mathbb{Y}_n) & \mapsto & \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n) \end{array} \right.$$

is measurable.

3.1 Statistical Model for Mapper

In this section, we study the convergence of the Mapper for a general generative model and a class of filter functions. We first introduce the generative model and next we present different settings depending on the nature of the filter function.

Generative model. The set of observations $\mathbb{X}_n$ is assumed to be composed of n independent points $X_1, ..., X_n$ sampled from a probability distribution $\mathbb{P}$ in $\mathbb{R}^D$. The support of $\mathbb{P}$ is denoted $\mathcal{X}_{\mathbb{P}}$ and is assumed to be a compact submanifold of $\mathbb{R}^D$ with positive reach and positive convexity radius, as in the setting of Theorem 2.7. We also assume that $0 < \text{diam}(\mathcal{X}_{\mathbb{P}}) \leq L$. Next, the probability distribution $\mathbb{P}$ is assumed to be (a, b)-standard for some constants $a > 0$ and $b \geq 1$, that is for any Euclidean ball $B(x, t)$ centered on $x \in \mathcal{X}$ with radius t :

$$P\left(B(x,t)\right) \geq \min(1, at^b).$$

This assumption is popular in the literature about set estimation (see for instance Cuevas, 2009; Cuevas and Rodríguez-Casal, 2004). It is also widely used in the TDA literature (Chazal et al., 2015b; Fasy et al., 2014; Chazal et al., 2015a). For instance, when $b = D$, this assumption is satisfied when the distribution is absolutely continuous with respect to the Hausdorff measure on $\mathcal{X}_{\mathbb{P}}$. We introduce the set $\mathcal{P}_{a,b} = \mathcal{P}_{a,b,\kappa,\rho,L}$ which is composed of all the (a, b)-standard probability distributions for which the support $\mathcal{X}_{\mathbb{P}}$ is a compact submanifold of $\mathbb{R}^D$ with reach larger than κ, convexity radius larger than ρ and diameter less than L.

Filter functions in the statistical setting. The filter function $f : \mathcal{X}_{\mathbb{P}} \mapsto \mathbb{R}$ for the Reeb graph is assumed as before to be a Morse-type function. Two different settings have to be considered regarding how the filter function is defined. In the first setting, the same filter function is used to define the Reeb graph and the Mapper. The Mapper can be defined by taking the exact values of the filter function at the observation points $f(X_1), \ldots, f(X_n)$. Note that this does not mean that the function f is completely known since, in our framework, knowing f would imply to know its domain and thus $\mathcal{X}_{\mathbb{P}}$ would be known which is of course not the case in practice. This first setting is referred to as the *exact filter setting* in the following. It corresponds to the situations where the Mapper algorithm is used with coordinate functions for instance. In the second setting, the filter function used for the Mapper is not available and an estimation of this filter function has to be computed from the data. This second setting is referred to as the *inferred filter setting* in the following. It corresponds to PCA or Laplacian eigenfunctions, distance functions (such as the DTM), or regression and density estimators.

Risk of Mapper. We study, in various settings, the problem of inferring a Reeb graph using Mappers and we use the metric d_Δ to assess the performance of the Mapper, seen as an estimator of the Reeb graph:

$$\mathbb{E}\left[d_\Delta\left(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_{\mathbb{P}})\right)\right],$$

where M_n is computed with the exact filter f or the inferred filter $\hat{f}$, depending on the context.

3.2 Reeb graph inference with exact filter and known generative model

We first consider the exact filter setting in the simplest situation where the parameters a and b of the generative model are known. In this setting, for given Rips parameter δ, gain g and resolution r, the Mapper $\mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$ is computed with $\mathbb{Y}_n = f(\mathbb{X}_n)$. We now tune the triple of parameters (r, g, δ) depending on the parameters a and b. Let $V_n(\delta_n) = \max\{|f(X) - f(X')| : X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta_n\}$. We choose for g a fixed value in $\left(\frac{1}{3}, \frac{1}{2}\right)$ and we take:

$$\delta_n = 8\left(\frac{2\log(n)}{an}\right)^{1/b} \quad \text{and} \quad r_n = \frac{V_n(\delta_n)}{g}.$$

We give below a general upper bound on the risk of M_n with these parameters, which depends on the regularity of the filter function and on the parameters of the generative model. We show a uniform convergence over a class of possible filter functions. This class of filters necessarily depends on the support of $\mathbb{P}$, so we define the class of filters for each probability measure in $\mathcal{P}_{a,b}$. For any $\mathbb{P} \in \mathcal{P}_{a,b}$, we let $\mathcal{F}(\mathbb{P}, \omega) = \mathcal{F}(\mathbb{P}, \omega, \underline{e})$ denote the set of filter functions $f : \mathcal{X}_{\mathbb{P}} \to \mathbb{R}$ such that f is Morse-type on $\mathcal{X}_{\mathbb{P}}$ with $\omega_f \leq \omega$ and such that $e_{\min}(\mathcal{X}_{\mathbb{P}}, f) > \underline{e}$.

Proposition 3.2. *Let ω be a modulus of continuity of f such that $\omega(x)/x$ is a non-increasing function on $\mathbb{R}^+$. For n large enough, the Mapper computed with parameters (r_n, g, δ_n) defined before satisfies*

$$\sup_{\mathbb{P} \in \mathcal{P}_{a,b}} \mathbb{E}\left[\sup_{f \in \mathcal{F}(\mathbb{P},\omega)} d_{\Delta}\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n\right)\right] \leq C\, \omega(\delta_n)$$

where the constant C only depends on a, b, and the geometric parameters of the model.

Assuming that $\omega(x)/x$ is non-increasing is not a very strong assumption. This property is satisfied in particular for concave modulus of functions. Thus, one can consider the concave majorant of ω_f, when it is finite (see for instance Section 6 in DeVore and Lorentz (1993)). As expected, we see that the rate of convergence of the Mapper to the Reeb graph directly depends on the regularity of the filter function and on the parameter b which roughly represents the intrinsic dimension of the data. For Lipschitz filter functions, the rate is similar to the one for persistence diagram inference Chazal et al. (2015b), namely it corresponds to the one of support estimation for the Hausdorff metric (see for instance Cuevas and Rodríguez-Casal (2004)) and Genovese et al. (2012a)). In the other cases where the filters only admit a concave modulus of continuity, we see that the "distortion" created by the filter function slows down the convergence of the Mapper to the Reeb graph.

We now give a lower bound that matches with the upper bound of Proposition 3.2.

Proposition 3.3. *Let ω be a modulus of continuity of f. Then, for any estimator $\hat{\mathrm{R}}_n$ of $\mathrm{R}_f(\mathcal{X}_{\mathbb{P}})$, we have*

$$\sup_{\mathbb{P} \in \mathcal{P}_{a,b}} \mathbb{E}\left[\sup_{f \in \mathcal{F}(\mathbb{P},\omega)} d_{\Delta}\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \hat{\mathrm{R}}_n\right)\right] \geq C\, \omega\left((an)^{-1/b}\right),$$

where the constant C only depends on a, b and the geometric parameters of the model.

Propositions 3.2 and 3.3 together show that, with the choice of parameters given before, M_n is minimax optimal up to a logarithmic factor $\log(n)$ inside the modulus of continuity. Note that the lower bound is also valid whether or not the coefficients a and b and the filter function f and its modulus of continuity are given.

3.3 Reeb graph inference with exact filter and unknown generative model

We still assume that the exact values $\mathbb{Y}_n = f(\mathbb{X}_n)$ of the filter on the point could can be computed and that at least an upper bound on the modulus of continuity of the filter is known. However, the parameters a and b are not assumed to be known anymore. We adapt a subsampling approach proposed by Fasy et al. (2014). As before, for given Rips parameter δ, gain g and resolution r, the Mapper $\mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$ is computed with $\mathbb{Y}_n = f(\mathbb{X}_n)$.

We introduce the sequence $s_n := \frac{n}{(\log n)^{1+\beta}}$ for some fixed value $\beta > 0$. Let $\hat{\mathbb{X}}_n^{s_n}$ be an arbitrary subset of $\mathbb{X}_n$ that contains s_n points. We tune the triple of parameters (r, g, δ) as follows: we choose for g a fixed value in $\left(\frac{1}{3}, \frac{1}{2}\right)$ and we take:

$$\delta_n = d_{\mathrm{H}}(\hat{\mathbb{X}}_n^{s_n}, \mathbb{X}_n) \quad \text{and} \quad r_n = \frac{V_n(\delta_n)}{g}, \tag{6}$$

where V_n is defined as in Section 3.2.

Proposition 3.4. *Let ω be a modulus of continuity of f such that $x \mapsto \omega(x)/x$ is a non-increasing function. Then, using the same notations as in the previous section, the Mapper M_n computed with parameters (r_n, g, δ_n) defined before satisfies*

$$\sup_{\mathbb{P} \in \mathcal{P}_{a,b}} \mathbb{E}\left[\sup_{f \in \mathcal{F}(\mathbb{P},\omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_\mathbb{P}), \mathrm{M}_n\right)\right] \leq C\, \omega\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b},$$

where the constant C only depends on a, b, and the geometric parameters of the model.

Up to logarithmic factors inside the modulus of continuity, we find that this Mapper is still minimax optimal over the class $\mathcal{P}_{a,b}$ by Proposition 3.3.

3.4 Reeb graph inference with inferred filter and unknown generative model

One of the nice properties of Mapper is that it can easily be computed with any filter function, including estimated filter functions such as PCA eigenfunctions, eccentricity functions, DTM functions, Laplacian eigenfunctions, density estimators, regression estimators, and many other filters directly estimated from the data. In this section, we assume that the *true filter* f is unknown but can be estimated from the data using an estimator $\hat{f}$. As before, parameters a and b are not assumed to be known and we have to tune the triple of parameters (r_n, g, δ_n).

In this context, the quantity V_n cannot be computed as before because there is no direct access to the values of f: we only know an estimation $\hat{f}$ of it. However, in many cases, an upper bound on the modulus of continuity of f is known, which makes possible the tuning of the parameters. For instance, PCA (and kernel) projectors, eccentricity functions, DTM functions (see Chazal et al. (2011)) are all 1-Lipschitz functions, and Corollary 3.5 below can be applied.

Let $\hat{V}_n(\delta_n) = \max\{|\hat{f}(X) - \hat{f}(X')| : X, X' \in \mathbb{X}_n, \|X - X'\| \leq \delta_n\}$, and let ω_1 be a modulus of continuity of f. Let

$$r_n := \frac{\max\{\omega_1(\delta_n), \hat{V}_n(\delta_n)\}}{g}. \tag{7}$$

Following the lines of the proof of Proposition 3.4 and applying Corollary 2.9, we obtain:

Corollary 3.5. *Let $f : \mathcal{X} \to \mathbb{R}$ be a Morse-type filter function and let $\hat{f} : \mathcal{X} \to \mathbb{R}$ be a Morse-type estimator of f. Assume that ω_1 (resp. ω_2) is a modulus of continuity of f (resp. $\hat{f}$). Let $\omega := \max\{\omega_1, \omega_2\}$ such that $x \mapsto \omega(x)/x$ is a non-increasing function. Let also $\hat{\mathrm{M}}_n := \mathrm{M}_{r_n, g, \delta_n}(\mathbb{X}_n, \hat{f}(\mathbb{X}_n))$ be the Mapper built on $\mathbb{X}_n$ with function $\hat{f}$ and parameters g, δ_n as in Equation (6), and r_n as in Equation (7). Then, $\hat{\mathrm{M}}_n$ statisfies*

$$\mathbb{E}\left[d_\Delta\left(\mathrm{R}_f(\mathcal{X}_\mathbb{P}), \hat{\mathrm{M}}_n\right)\right] \leq C\omega\left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b} + \mathbb{E}\left[\max_{1 \leq i \leq n} |f(X_i) - \hat{f}(X_i)|\right],$$

where C only depends on a, b, and the geometric parameters of the model.

PCA eigenfunctions. In the setting of this article, the measure μ has a finite second moment. Following Biau and Mas (2012), we define the covariance operator $\Gamma(\cdot) := \mathbb{E}(\langle X, \cdot \rangle X)$ and we let Π_k denote the orthogonal projection onto the space spanned by the k-th eigenvector of Γ. In practice, we consider the empirical version of the covariance operator

$$\hat{\Gamma}_n(\cdot) := \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \cdot \rangle X_i$$

and the empirical projection $\hat{\Pi}_k$ onto the space spanned by the k-th eigenvector of $\hat{\Gamma}_n$. According to Biau and Mas (2012)(see also Blanchard et al. (2007); Shawe-Taylor et al. (2005)), we have

$$\mathbb{E}\left[\|\Pi_k - \hat{\Pi}_k\|_\infty\right] = O\left(\frac{1}{\sqrt{n}}\right).$$

This, together with Corollary 3.5 and the fact that both Π_k and $\hat{\Pi}_k$ are 1-Lipschitz, gives that the rate of convergence of the Mapper of $\hat{\Pi}_k(\mathbb{X}_n)$ computed with parameters δ_n and g as in Equation (6), and r_n as in Equation (7), i.e. $r_n := g^{-1}\delta_n$, satisfies

$$\mathbb{E}\left[d_\Delta\left(\mathrm{R}_{\Pi_k}(\mathcal{X}_\mathbb{P}), \mathrm{M}_{r_n,g,\delta_n}(\mathbb{X}_n, \hat{\Pi}_k(\mathbb{X}_n))\right)\right] \lesssim \left(\frac{\log(n)^{2+\beta}}{n}\right)^{1/b} \vee \frac{1}{\sqrt{n}}.$$

Hence, the rate of convergence of Mapper is not deteriorated by using $\hat{\Pi}_k$ instead of Π_k if the intrinsic dimension b of the support of $\mathcal{X}$ is at least 2.

The distance to measure. It is well known that TDA methods may fail completely in the presence of outliers. To address this issue, Chazal et al. (2011) introduced an alternative distance function which is robust to noise, the *distance-to-a-measure* (DTM). A similar analysis as with the PCA filter can be carried out with the DTM filter using the rates of convergence proven in Chazal et al. (2016b).

4 Confidence sets for Reeb signatures

4.1 Confidence sets for extended persistence diagrams

In practice, computing a Mapper M_n and its signature $\mathrm{Dg}(\mathrm{M}_n, f_\mathcal{I})$ is not sufficient: we need to know how accurate these estimations are. One natural way to answer this problem is to provide a confidence set for the Mapper using the bottleneck distance. For $\alpha \in (0,1)$, we look for some value $\eta_{n,\alpha}$ such that

$$P\left(d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_\mathbb{P})) \geq \eta_{n,\alpha}\right) \leq \alpha$$

or at least such that

$$\limsup_{n\to\infty} P\left(d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_\mathbb{P})) \geq \eta_{n,\alpha}\right) \leq \alpha.$$

Let

$$\mathcal{M}_\alpha := \{\mathrm{R} \in \mathcal{R} \,|\, d_\Delta(\mathrm{M}_n, \mathrm{R}) \leq \alpha\}$$

be the closed ball of radius α in the bottleneck distance and centered at the Mapper M_n in the space of Reeb graphs $\mathcal{R}$. Following Fasy et al. (2014), we can visualize the signatures of the points

belonging to this ball in various ways. One first option is to center a box of side length 2α at each point of the extended persistence diagram of M_n—see the right columns of Figure 5 and Figure 6 for instance. An alternative solution is to visualize the confidence set by adding a band at (vertical) distance $\eta_{n,\alpha}/2$ from the diagonal (the bottleneck distance being defined for the ℓ_∞ norm). The points outside the band are then considered as significant topological features, see Fasy et al. (2014) for more details.

Several methods have been proposed in Fasy et al. (2014) and Chazal et al. (2014) to define confidence sets for persistence diagrams. We now adapt these ideas to provide confidence sets for Mappers. Except for the bottleneck bootstrap (see further), all the methods proposed in these two articles rely on the stability results for persistence diagrams, which say that persistence diagrams equipped with the bottleneck distance are stable under Hausdorff or Wasserstein perturbations of the data. Confidence sets for diagrams are then directly derived from confidence sets in the sample space. Here, we follow a similar strategy using Theorem 2.7, as explained in the next section.

4.2 Confidence sets derived from Theorem 2.7

In this section, we always assume that an upper bound ω on the exact modulus of continuity ω_f of the filter function is known. We start with the following remark: if we can take δ of the order of $d_{\mathrm{H}}(\mathcal{X}_{\mathbb{P}}, \mathbb{X}_n)$ in Theorem 2.7 and if all the conditions of the theorem are satisfied, then $d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_{\mathbb{P}}))$ can be bounded in terms of $\omega(d_{\mathrm{H}}(\mathcal{X}_{\mathbb{P}}, \mathbb{X}_n))$. This means that we can adapt the methods of Fasy et al. (2014) to Mappers.

Known generative model. Let us first consider the simplest situation where the parameters a an b are also known. Following Section 3.2, we choose for g a fixed value in $\left(\frac{1}{3}, \frac{1}{2}\right)$ and we take

$$\delta_n = 8\left(\frac{2\log(n)}{an}\right)^{1/b} \quad \text{and} \quad r_n = \frac{V_n(\delta_n)}{g},$$

where V_n is defined as in Section 3.2. Let $\varepsilon_n = d_{\mathrm{H}}(\mathcal{X}_{\mathbb{P}}, \mathbb{X}_n)$. As shown in the proof of Proposition 3.2 (see Appendix A.5), for n large enough, Assumption (1) and (2) are always satisfied and then

$$P\left(d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_{\mathbb{P}})) \geq \eta\right) \leq P\left(\delta_n \geq \omega^{-1}\left(\frac{\eta}{(2g)^{-1}+2}\right)\right).$$

Consequently,

$$\begin{aligned} P\left(d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_{\mathbb{P}})) \geq \eta\right) &\leq P\left(d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_{\mathbb{P}})) \geq \eta \cap \varepsilon_n \leq 4\delta_n\right) + P\left(\varepsilon_n > 4\delta_n\right) \\ &\leq \mathbb{I}_{\omega(\delta_n) \geq \frac{2g}{1+4g}\eta} + \min\left\{1, \frac{2^b}{2\log(n)n}\right\} \\ &=: \Phi_n(\eta). \end{aligned}$$

where Φ_n depends on the parameters of the model (or some bounds on these parameters) which are here assumed to be known. Hence, given a probability level α, one has:

$$P\left(d_\Delta(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X}_{\mathbb{P}})) \geq \Phi_n^{-1}(\alpha)\right) \leq \alpha.$$

Unknown generative model. We now assume that a and b are unknown. To compute confidence sets for the Mapper in this context, we approximate the distribution of $d_{\mathrm{H}}(\mathcal{X}_{\mathbb{P}}, \mathbb{X}_n)$ using the distribution of $d_{\mathrm{H}}(\hat{\mathbb{X}}_n^{s_n}, \mathbb{X}_n)$ conditionally to $\mathbb{X}_n$. There are $N_1 = \binom{n}{s_n}$ subsets of size s_n inside $\mathbb{X}_n$, so we let $\mathbb{X}_{s_n}^1, \ldots, \mathbb{X}_{s_n}^{N_1}$ denote all the possible configurations. Define

$$L_n(t) = \frac{1}{N_1} \sum_{k=1}^{N_1} \mathbb{I}_{d_{\mathrm{H}}(\mathbb{X}_{s_n}^k, \mathbb{X}_n) > t}.$$

Let s be the function on $\mathbb{N}$ defined by $s(n) = s_n$ and let $s_n^2 := s(s(n))$. There are $N_2 = \binom{n}{s_n^2}$ subsets of size s_n^2 inside $\mathbb{X}_n$. Again, we let $\mathbb{X}_{s_n^2}^k$, $1 \leq k \leq N_2$, denote these configurations and we also introduce

$$F_n(t) = \frac{1}{N_2} \sum_{k=1}^{N_2} \mathbb{I}_{d_{\mathrm{H}}\left(\mathbb{X}_{s_n^2}^k, \mathbb{X}_{s_n}\right) > t}.$$

Proposition 4.1. *Let $\eta > 0$. Then, one has the following confidence set:*

$$P\left(d_\Delta(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n) \geq \eta\right) \leq F_n\left(\frac{1}{4}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right) + L_n\left(\frac{1}{4}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right) + o\left(\frac{s_n}{n}\right)^{1/4}.$$

Both F_n and L_n can be computed in practice, or at least approximated using Monte Carlo procedures. The upper bound on $P(d_\Delta(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n) \geq \eta)$ then provides an asymptotic confidence region for the persistence diagram of the Mapper M_n, which can be explicitly computed in practice. See the green squares in the first row of Figure 5. The main drawback of this approach is that it requires to know an upper bound on the modulus of continuity ω and, more importantly, the number of observations has to be very large, which is not the case on our examples in Section 5.

Modulus of continuity of the filter function. As shown in Proposition 4.1, the modulus of continuity of the filter function is a key quantity to describe the confidence regions. Inferring the modulus of continuity of the filter from the data is a tricky problem. Fortunately, in practice, even in the inferred filter setting, the modulus of continuity of the function is known in many situations. For instance, projections such as PCA eigenfunctions and DTM functions are 1-Lipschitz.

4.3 Bottleneck Bootstrap

The two methods given before both require an explicit upper bound on the modulus of continuity of the filter function. Moreover, these methods both rely on the approximation result Theorem 2.7, which often leads to conservative confidence sets. An alternative strategy is the bottleneck bootstrap introduced in Chazal et al. (2014), and which we now apply to our framework.

The bootstrap is a general method for estimating standard errors and computing confidence intervals. Let $\mathbb{P}_n$ be the empirical measure defined from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $(X_1^*, Y_1^*) \ldots, (X_n^*, Y_n^*)$ be a sample from $\mathbb{P}_n$ and let also M_n^* be the random Mapper defined from this sample. We then take for $\hat{\eta}_\alpha$ the quantity $\hat{\eta}_\alpha^*$ defined by

$$P(d_\Delta(\mathrm{M}_n^*, \mathrm{M}_n) > \hat{\eta}_\alpha^* \,|\, X_1, \ldots, X_n) = \alpha. \tag{8}$$

Note that $\hat{\eta}_\alpha^*$ can be easily estimated with Monte Carlo procedures. It has been shown in Chazal et al. (2014) that the bottleneck bootstrap is valid when computing the sublevel sets of a density

estimator. The validity of the bottleneck bootstrap has not been proven for the extended persistence diagram of any distance function. For Mapper, it would require to write $d_\Delta(\mathrm{M}_n^*, \mathrm{M}_n)$ in terms of the distance between the extrema of the filter function and the ones of the interpolation of the filter function on the Rips. We leave this problem open in this article.

Extension of the analysis. As pointed out in Section 2.1, many versions of the discrete Mapper exist in the literature. One of them, called the *edge-based* version $\mathrm{M}^{\Delta}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$, is described in Section 7 of (Carrière and Oudot (2015)). The main advantage of this edge-based version is that it allows for finer resolutions than the usual Mapper while remaining fast to compute. Our analysis can actually handle this edge-based version as well by replacing gr by r in Assumption (2) of Theorem 2.7—see Remark 2.8, and changing constants accordingly in the proofs. In particular, this improves the resolution r_n in Equation (6) since $g^{-1}V_n(\delta_n)$ becomes $V_n(\delta_n)$. Hence, we use this edge-based version in Section 5, where this improvement on the resolution r_n allows us to compensate for the low number of observations in some datasets.

5 Numerical experiments

In this section, we provide few examples of parameter selections and confidence regions (which are union of squares in the extended persistence diagrams) obtained with bottleneck bootstrap. The interpretation of these regions is that squares that intersect the diagonal, which are drawn in pink color, represent topological features in the Mappers that may be horizontal or artifacts due to the cover, and that may not be present in the Reeb graph. We show in Figure 5 various Mappers (in each node of the Mappers, the left number is the cluster ID and the right number is the number of observations in that cluster) and 85 percent confidence regions computed on various datasets. All δ parameters and resolutions were computed with Equation (6) (the δ parameters were also averaged over $N = 100$ subsamplings with $\beta = 0.001$), and all gains were set to 40%. The code we used is expected to be added in the next release of The GUDHI Project (2015), and should then be available soon. The confidence regions were computed by bootstrapping data 100 times. Note that computing confidence regions with Proposition 4.1 is possible, but the numbers of observations in all of our datasets were too low, leading to conservative confidence regions that did not allow for interpretation.

5.1 Mappers and confidence regions

Synthetic example. We computed the Mapper of an embedding of the Klein bottle into $\mathbb{R}^4$ with 10,000 points with the height function. In order to illustrate the conservativity of confidence regions computed with Proposition 4.1, we also plot these regions for an embedding with 10,000,000 points using the fact that the height function is 1-Lipschitz. Corresponding squares are drawn in green color. Their very large sizes show that Proposition 4.1 requires a very large number of observations in practice. See the first row of Figure 5.

3D shapes. We computed the Mapper of an ant shape and a human shape from Chen et al. (2009) embedded in $\mathbb{R}^3$ (with 4,706 and 6,370 points respectively) Both Mappers were computed with the height function. One can see that the confidence squares for the features that are almost

horizontal (such as the small branches in the Mapper of the ant) intersect indeed the diagonal. See the second and third rows of Figure 5.

Miller-Reaven dataset. The first dataset comes from the Miller-Reaven diabetes study that contains 145 observations of patients suffering or not from diabete. Observations were mapped into $\mathbb{R}^5$ by computing various medical features. Data can be obtained in the "locfit" R-package. In Reaven and Miller (1979), the authors identified two groups of diseases with the projection pursuit method, and in Singh et al. (2007), the authors applied Mapper with hand-crafted parameters to get back this result. Here, we normalized the data to zero mean and unit variance, and we obtained the two flares in the Mapper computed with the eccentricity function. Moreover, these flares are at least 85 percent sure since the confidence squares on the corresponding points in the extended persistence diagrams do not intersect the diagonal. See the first row of Figure 6.

COIL dataset. The second dataset is an instance of the 16,384-dimensional COIL dataset Nene et al. (1996). It contains 72 observations, each of which being a picture of a duck taken at a specific angle. Despite the low number of observations and the large number of dimensions, we managed to retrieve the intrinsic loop lying in the data using the first PCA eigenfunction. However, the low number of observations made the bootstrap fail since the confidence squares computed around the points that represent this loop in the extended persistence diagram intersect the diagonal. See the second row of Figure 6.

5.2 Noisy data

Denoising Mapper. An important drawback of Mapper is its sensitivity to noise and outliers. See the crater dataset in Figure 7, for instance. Several answers have been proposed for recovering the correct persistence homology from noisy data. The idea is to use an alternative filtration of simplical compexes instead of the Rips filtration. A first option is to consider the upper level sets of a density estimator rather then the distance to the sample (see Section 4.4 in Fasy et al. (2014)). Another solution is to consider the sublevel sets of the DTM and apply persistence homology inference in Chazal et al. (2014).

Crater dataset. To handle noise in our crater dataset, we simply smoothed the dataset by computing the empirical DTM with 10 neighbors on each point and removing all points with DTM less than 40 percent of the maximum DTM in the dataset. Then we computed the Mapper with the height function. One can see that all topological features in the Mapper that are most likely artifacts due to noise (like the small loops and connected components) have corresponding confidence squares that intersect the diagonal in the extended persistence diagram. See Figure 7.

6 Conclusion

In this article, we provided a statistical analysis of the Mapper. Namely, we proved the fact that the Mapper is a measurable construction in Proposition 3.1, and we used the approximation Theorem 2.7 to show that the Mapper is a minimax optimal estimator of the Reeb graph in various contexts—see Propositions 3.2, 3.3 and 3.4—and that corresponding confidence regions can be computed—see Proposition 4.1 and Section 4.3. Along the way, we derived rules of thumb to

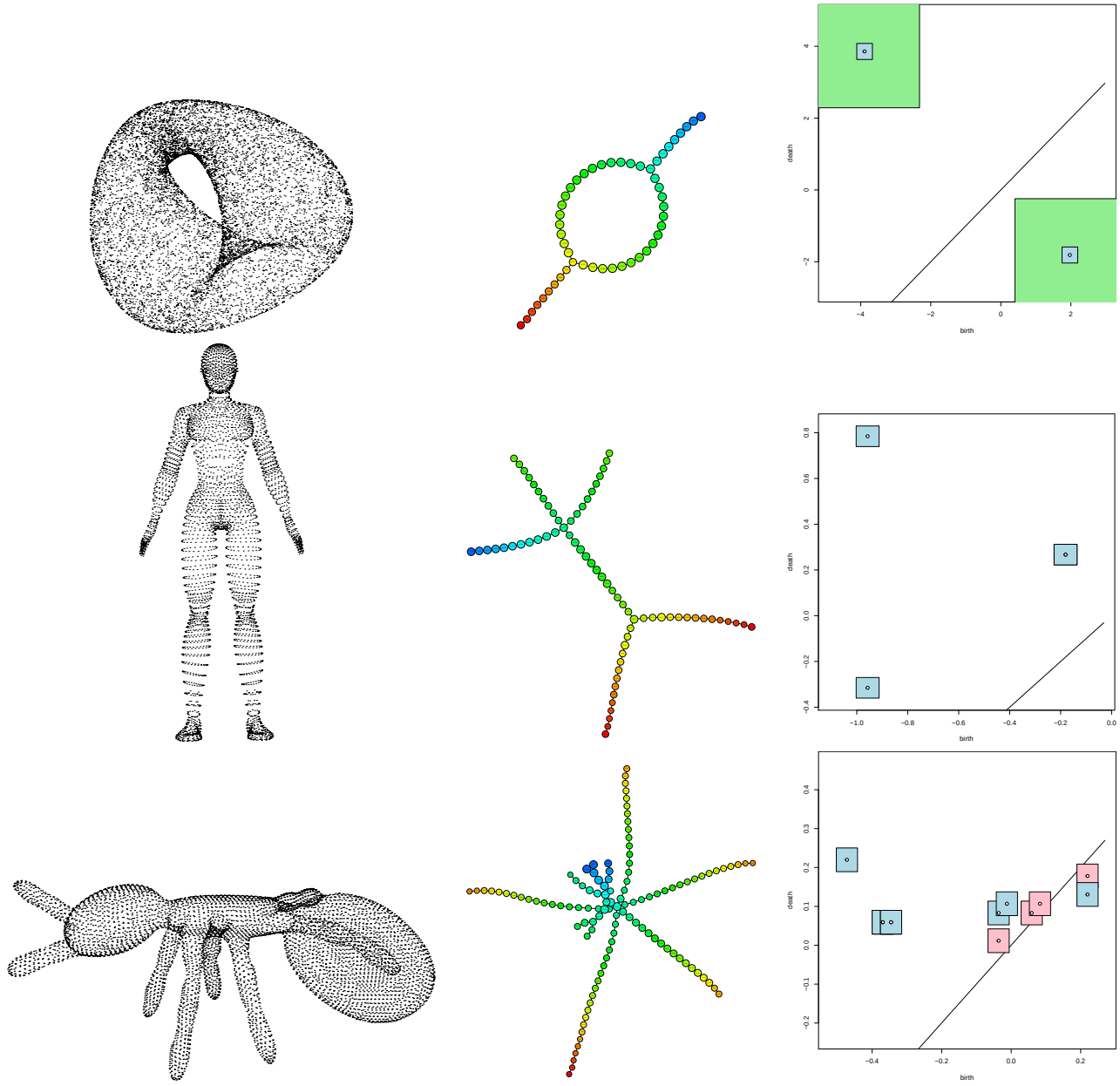

Figure 5: Mappers computed with automatic tuning (middle) and 85 percent confidence regions for their topological features (right) are provided for an embedding of the Klein Bottle into $\mathbb{R}^4$ (first row), a 3D human shape (second row) and a 3D ant shape (third row).

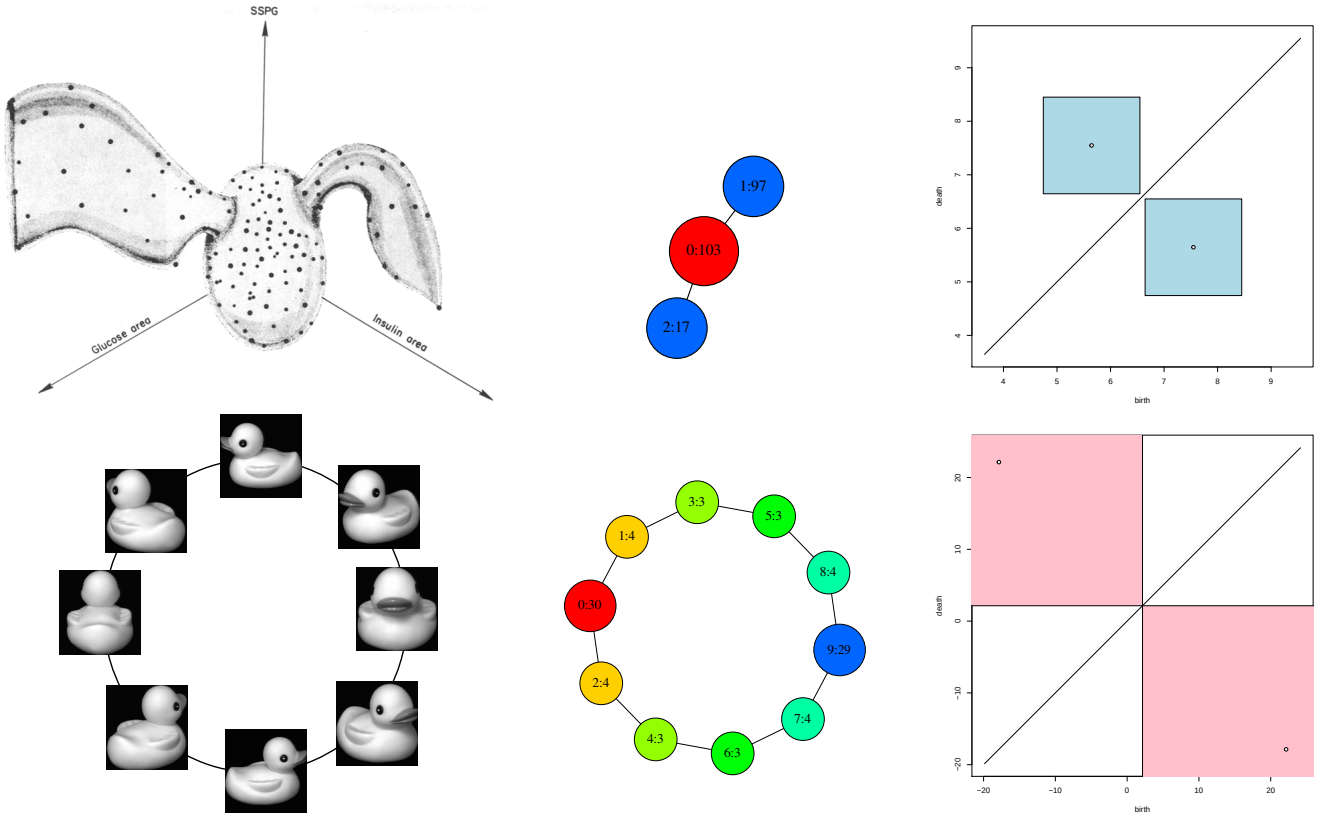

Figure 6: Mappers computed with automatic tuning (middle) and 85 percent confidence regions for their topological features (right) are provided for the Reaven-Miller dataset (first row) and the COIL dataset (second row).

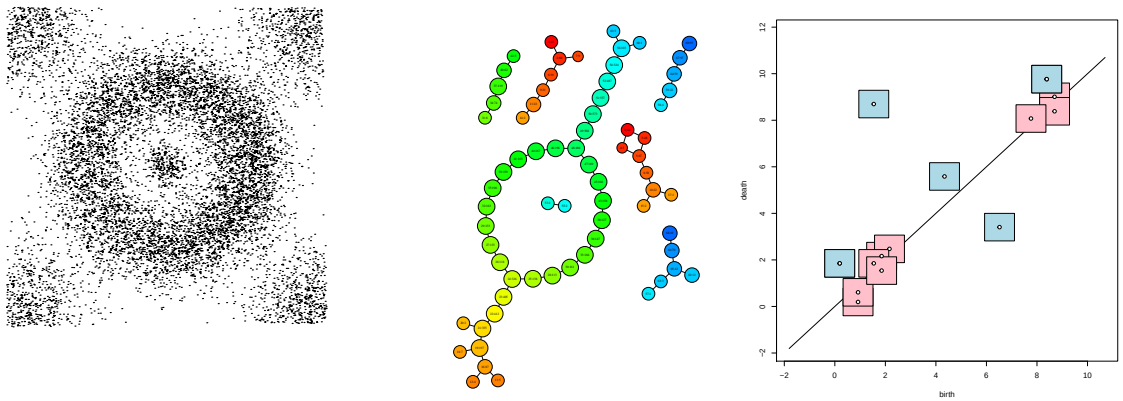

Figure 7: Mappers computed with automatic tuning (middle) and 85 percent confidence regions for their topological features (right) are provided for a a noisy crater in the Euclidean plane.

automatically tune the parameters of the Mapper with Equation (6). Finally, we provided few examples of our methods on various datasets in Section 5.

Future directions. We plan to investigate several questions for future work.

- We will work on adapting results from Chazal et al. (2014) to prove the validity of bootstrap methods for computing confidence regions on the Mapper, since we only used bootstrap methods empirically in this article.
- We believe that using weighted Rips complexes Buchet et al. (2015) instead of the usual Rips complexes would improve the quality of the confidence regions on the Mapper features, and would probably be a better way to deal with noise that our current solution.
- We plan to adapt our statistical setting to the question of selecting variables, which is one of the main applications of the Mapper in practice.

A Proofs

A.1 Preliminary results

In order to prove the results of this article, we need to state several preliminary definitions and theorems. All of them can be found, together with their proofs, in Dey and Wang (2013) and Carrière and Oudot (2015). In this section, we let $\mathbb{X}_n \subset \mathcal{X}$ be a point cloud of n points sampled on a submanifold $\mathcal{X}$ embedded in $\mathbb{R}^D$, with positive reach rch and convexity radius ρ. Let $f : \mathcal{X} \to \mathbb{R}$ be a Morse-type filter function such that $e_{\min}(\mathcal{X}, f) > 0$, $\mathcal{I}$ be a minimal open cover of the range of f with resolution r and gain g, $|\mathrm{Rips}_\delta(\mathbb{X}_n)|$ denote a geometric realization of the Rips complex built on top of $\mathbb{X}_n$ with parameter δ, and $f^{\mathrm{PL}} : |\mathrm{Rips}_\delta(\mathbb{X}_n)| \to \mathbb{R}$ be the piecewise-linear interpolation of f on the simplices of $\mathrm{Rips}_\delta(\mathbb{X}_n)$.

Definition A.1. *Let $G = (\mathbb{X}_n, E)$ be a graph built on top of $\mathbb{X}_n$. Let $e = (X, X') \in E$ be an edge of G, and let $I(e)$ be the open interval $(\min\{f(X), f(X')\}, \max\{f(X), f(X')\})$. Then e is said to be* intersection-crossing *if there is a pair of consecutive intervals $I, J \in \mathcal{I}$ such that $\emptyset \neq I \cap J \subseteq I(e)$.*

Theorem A.2. *Let $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$ denote the 1-skeleton of $\mathrm{Rips}_\delta(\mathbb{X}_n)$. If $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$ has no intersection-crossing edges, then $\mathrm{M}_{r,g,\delta}(\mathbb{X}_n, f(\mathbb{X}_n))$ and $\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$ are isomorphic as combinatorial graphs.*

Theorem A.3. *Let $f : \mathcal{X} \to \mathbb{R}$ be a Morse-type function. Then, we have the following inequality between extended persistence diagrams:*

$$d_\Delta(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}}), \mathrm{Dg}(\mathrm{M}_{r,g}(\mathcal{X}, f), f_{\mathcal{I}})) \leq \frac{r}{2}. \tag{9}$$

Moreover, given another Morse-type function $\hat{f} : \mathcal{X} \to \mathbb{R}$, we have the following inequality:

$$d_\Delta(\mathrm{Dg}(\mathrm{M}_{r,g}(\mathcal{X}, f), f_{\mathcal{I}}), \mathrm{Dg}(\mathrm{M}_{r,g}(\mathcal{X}, \hat{f}), \hat{f}_{\mathcal{I}})) \leq r + \|f - \hat{f}\|_\infty. \tag{10}$$

Theorem A.4. *If $4d_{\mathrm{H}}(\mathcal{X}, \mathbb{X}_n) \leq \delta \leq \min\{rch/4, \rho/4\}$ and $\omega(\delta) \leq e_{\min}/2$, then*

$$d_\Delta(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}}), \mathrm{Dg}(\mathrm{R}_{f^{\mathrm{PL}}}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|), f^{\mathrm{PL}}_{\mathrm{R}})) \leq 2\omega(\delta).$$

Note that the original version of this theorem is only proven for Lipschitz functions in Dey and Wang (2013), but it extends at no cost to functions with modulus of continuity.

A.2 Proof of Theorem 2.7

Let $|\mathrm{Rips}_\delta(\mathbb{X}_n)|$ denote a geometric realization of the Rips complex built on top of $\mathbb{X}_n$ with parameter δ. Moreover, let $f^{\mathrm{PL}} : |\mathrm{Rips}_\delta(\mathbb{X}_n)| \to \mathbb{R}$ be the piecewise-linear interpolation of f on the simplices of $\mathrm{Rips}_\delta(\mathbb{X}_n)$, whose 1-skeleton is denoted by $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$. Since $(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$ is a metric space, we also consider its Reeb graph $\mathrm{R}_{f^{\mathrm{PL}}}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|)$, with induced function $f^{\mathrm{PL}}_{\mathrm{R}}$, and its Mapper $\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$, with induced function $f^{\mathrm{PL}}_{\mathcal{I}}$. See Figure 8. Then, the following inequalities lead to the result:

$$
\begin{aligned}
d_\Delta(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n) &= d_\Delta\left(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}}), \mathrm{Dg}(\mathrm{M}_n, f_{\mathcal{I}})\right) \\
&= d_\Delta\left(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}}), \mathrm{Dg}(\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}}), f^{\mathrm{PL}}_{\mathcal{I}})\right) && (11)\\
&\leq d_\Delta\left(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}}), \mathrm{Dg}(\mathrm{R}_{f^{\mathrm{PL}}}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|), f^{\mathrm{PL}}_{\mathrm{R}})\right) \\
&\quad + d_\Delta\left(\mathrm{Dg}(\mathrm{R}_{f^{\mathrm{PL}}}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|), f^{\mathrm{PL}}_{\mathrm{R}}), \mathrm{Dg}(\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}}), f^{\mathrm{PL}}_{\mathcal{I}})\right) && (12)\\
&\leq 2\omega(\delta) + \frac{r}{2}. && (13)
\end{aligned}
$$

Let us prove every (in)equality:

Equality (11). Let $X_1, X_2 \in \mathbb{X}_n$ such that (X_1, X_2) is an edge of $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$ i.e. $\|X_1 - X_2\| \leq \delta$. Then, according to (2): $|f(X_1) - f(X_2)| \leq gr$. Hence, there is no $s \in \{1, \ldots, S-1\}$ such that $I_s \cap I_{s+1} \subseteq [\min\{f(X_1), f(X_2)\}, \max\{f(X_1), f(X_2)\}]$. It follows that there are no intersection-crossing edges in $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$. Then, according to Theorem A.2, there is a graph isomorphism $i : \mathrm{M}_n = \mathrm{M}_{r,g,\delta}(\mathbb{X}_n, f(\mathbb{Y}_n)) \to \mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$. Since $f_{\mathcal{I}} = f^{\mathrm{PL}}_{\mathcal{I}} \circ i$ by definition of $f_{\mathcal{I}}$ and $f^{\mathrm{PL}}_{\mathcal{I}}$, the equality follows.

Inequality (12). This inequality is just an application of the triangle inequality.

Inequality (13). According to (1), we have $\omega(\delta) \leq e_{\min}/2$ and $\delta \leq \min\{rch/4, \rho/4\}$. Aaccording to (3), we also have $\delta \geq 4d_{\mathrm{H}}(\mathcal{X}, \mathbb{X}_n)$. Hence, we have

$$d_\Delta(\mathrm{Dg}(\mathrm{R}_f(\mathcal{X}), f_{\mathrm{R}}), \mathrm{Dg}(\mathrm{R}_{f^{\mathrm{PL}}}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|), f^{\mathrm{PL}}_{\mathrm{R}})) \leq 2\omega(\delta),$$

according to Theorem A.4. Moreover, we have

$$d_\Delta(\mathrm{Dg}(\mathrm{R}_{f^{\mathrm{PL}}}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|), f^{\mathrm{PL}}_{\mathrm{R}}), \mathrm{Dg}(\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}}), f^{\mathrm{PL}}_{\mathcal{I}})) \leq \frac{r}{2},$$

according to Equation (9).

A.3 Proof of Corollary 2.9

Let $|\mathrm{Rips}_\delta(\mathbb{X}_n)|$ denote a geometric realization of the Rips complex built on top of $\mathbb{X}_n$ with parameter δ. Moreover, let $f^{\mathrm{PL}} : |\mathrm{Rips}_\delta(\mathbb{X}_n)| \to \mathbb{R}$ be the piecewise-linear interpolation of f on the simplices of $\mathrm{Rips}_\delta(\mathbb{X}_n)$, whose 1-skeleton is denoted by $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$. Similarly, let $\hat{f}^{\mathrm{PL}}$ be the piecewise-linear interpolation of $\hat{f}$ on the simplices of $\mathrm{Rips}^1_\delta(\mathbb{X}_n)$. As before, since $(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$ and $(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, \hat{f}^{\mathrm{PL}})$ are metric spaces, we also consider their Mappers $\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$ and $\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, \hat{f}^{\mathrm{PL}})$. Then, the following inequalities lead to the result:

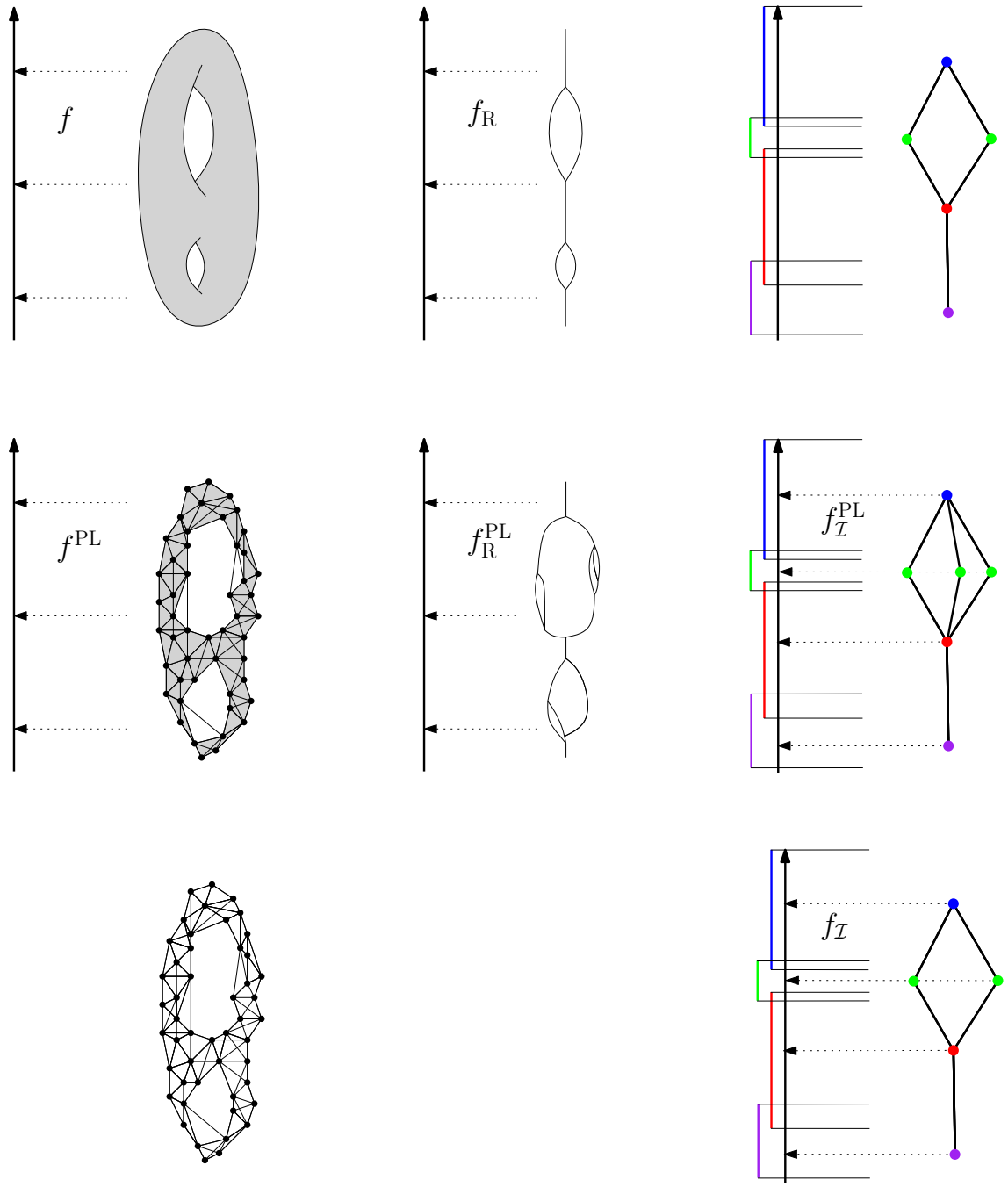

Figure 8: Examples of the function defined on the original space (left column), its induced function defined on the Reeb graph (middle column) and the function defined on the Mapper (right column). Note that the Mapper computed from the geometric realization of the Rips complex (middle row, right) is not isomorphic to the standard Mapper (third row, right), since there is an intersection-crossing edge in the neighborhood graph.

$$
\begin{aligned}
d_\Delta(\mathrm{R}_f(\mathcal{X}), \hat{\mathrm{M}}_n) &\leq d_\Delta(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n) + d_\Delta(\mathrm{M}_n, \hat{\mathrm{M}}_n) \text{ by the triangle inequality} \\
&= d_\Delta(\mathrm{R}_f(\mathcal{X}), \mathrm{M}_n) + d_\Delta(\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}}), \mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, \hat{f}^{\mathrm{PL}})) \qquad (14) \\
&\leq \frac{r}{2} + 2\omega(\delta) + d_\Delta(\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}}), \mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, \hat{f}^{\mathrm{PL}})) \text{ by Theorem 2.7} \\
&\leq \frac{r}{2} + 2\omega(\delta) + r + \|f^{\mathrm{PL}} - \hat{f}^{\mathrm{PL}}\|_\infty \text{ by Equation (10)} \\
&= \frac{3r}{2} + 2\omega(\delta) + \max\{|f(X) - \hat{f}(X)| \,:\, X \in \mathbb{X}_n\}
\end{aligned}
$$

Let us prove Equality (14). By definition of r, there are no intersection-crossing edges for both f and $\hat{f}$. According to Theorem A.2, $\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, f^{\mathrm{PL}})$ and M_n are isomorphic and similarly for $\mathrm{M}_{r,g}(|\mathrm{Rips}_\delta(\mathbb{X}_n)|, \hat{f}^{\mathrm{PL}})$ and $\hat{\mathrm{M}}_n$. See also the proof of Equality 11.

A.4 Proof of Proposition 3.1

We check that not only the topological signature of Mapper but also Mapper itself is a measurable object and thus can be seen as an estimator of a target Reeb graph. This problem is more complicated than for the statistical framework of persistence diagram inference, for which the existing stability results give for free that persistence estimators are measurable for adequate sigma algebras.

Let $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ denote the extended real line. Given a fixed integer $n \geq 1$, let $\mathcal{C}_{[n]}$ be the set of abstract simplicial complexes over a fixed set of n vertices. We see $\mathcal{C}_{[n]}$ as a subset of the power set $2^{2^{[n]}}$, where $[n] = \{1, \cdots, n\}$, and we implicitly identify $2^{[n]}$ with the set $[2^n]$ via the map assigning to each subset $\{i_1, \cdots, i_k\}$ the integer $1 + \sum_{j=1}^k 2^{i_j - 1}$. Given a fixed parameter $\delta > 0$, we define the application

$$
\Phi_1 : \left\{ \begin{array}{rcl} (\mathbb{R}^D)^n \times \mathbb{R}^n & \to & \mathcal{C}_{[n]} \times \bar{\mathbb{R}}^{2^n} \\ (\mathbb{X}_n, \mathbb{Y}_n) & \mapsto & (K, f) \end{array} \right.
$$

where K is the abstract Rips complex of parameter δ over the n labeled points in $\mathbb{R}^D$, minus the intersection-crossing edges and their cofaces, and where f is a function defined on the simplices of K by

$$
f : \left\{ \begin{array}{rcl} 2^n & \to & \bar{\mathbb{R}} \\ \sigma & \mapsto & \left\{ \begin{array}{ll} \max_{i \in \sigma} \mathbb{Y}_i & \text{if } \sigma \in K \\ +\infty & \text{otherwise.} \end{array} \right. \end{array} \right.
$$

The space $(\mathbb{R}^D)^n \times \mathbb{R}^n$ is equipped with the standard topology, denoted by T_1, inherited from $\mathbb{R}^{(D+1)n}$. The space $\mathcal{C}_{[n]} \times \bar{\mathbb{R}}^{2^n}$ is equipped with the product of the discrete topology on $\mathcal{C}_{[n]}$ and the topology induced by the extended distance $d_\infty(f, g) := \max\{|f(\sigma) - g(\sigma)| \,:\, \sigma \in 2^n,\ f(\sigma) \text{ or } g(\sigma) \neq +\infty\}$ on $\bar{\mathbb{R}}^{2^n}$. This product is denoted by T_2 hereafter.

Note that the map $(\mathbb{X}_n, \mathbb{Y}_n) \mapsto K$ is piecewise-constant, with jumps located at the hypersurfaces defined by $\|X_i - X_j\|^2 = \delta^2$ (for combinatorial changes in the Rips complex) or $Y_i = \mathrm{cst} \in \mathrm{End}(\mathcal{I})$ (for changes in the set of intersection-crossing edges) in $(\mathbb{R}^D)^n \times \mathbb{R}^n$. We can then define a finite measurable partition $(\mathcal{D}_\ell)_{\ell \in L}$ of $(\mathbb{R}^D)^n \times \mathbb{R}^n$ whose boundaries are included in these hypersurfaces,

and such that $(\mathbb{X}_n, \mathbb{Y}_n) \mapsto K$ is constant over each set $\mathcal{D}_\ell$. As a byproduct, we have that $(\mathbb{X}_n, \mathbb{Y}_n) \mapsto f$ is continuous over each set $\mathcal{D}_\ell$.

We now define the operator

$$\Phi_2 : \left\{ \begin{array}{rcl} \mathcal{C}_{[n]} \times \bar{\mathbb{R}}^{2^n} & \to & \mathcal{A} \\ (K, f) & \mapsto & (|K|, f^{\mathrm{PL}}) \end{array} \right.$$

where $\mathcal{A}$ denotes the class of topological spaces filtered by Morse-type functions, and where f^{PL} is the piecewise-linear interpolation of f on the geometric realization $|K|$ of K. For a fixed simplicial complex K, the persistence diagrams of f and f^{PL} are identical—see e.g. Morozov (2008), therefore the map Φ_2 is distance-preserving (hence continuous) in the pseudometrics d_Δ on the domain and codomain. Since the topology T_2 on $\mathcal{C}_{[n]} \times \bar{\mathbb{R}}^{2^n}$ is a refinement[1] of the topology induced by d_Δ, the map Φ_2 is also continuous when $\mathcal{C}_{[n]} \times \bar{\mathbb{R}}^{2^n}$ is equipped with T_2.

Let now $\Phi_3 : \mathcal{A} \to \mathcal{R}$ map each Morse-type pair $(\mathcal{X}, f)$ to its Mapper $(\mathrm{M}_{r,g}(\mathcal{X}, f), f_{\mathcal{I}})$. Note that, similarly to Φ_1, the map Φ_3 is piecewise-constant, since combinatorial changes in $\mathrm{M}_{r,g}(\mathcal{X}, f)$ are located at the regions $\mathrm{Crit}(f) \cap \mathrm{End}(\mathcal{I}) \neq \emptyset$, and since $f_{\mathcal{I}}$ depends only on the combinatorial structure of $\mathrm{M}_{r,g}(\mathcal{X}, f)$. Hence, Φ_3 is measurable in the pseudometric d_Δ. Moreover, $\mathrm{M}_{r,g}(|K|, f^{\mathrm{PL}})$ is isomorphic to $\mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$ by Theorem A.2 since all intersection-crossing edges were removed in the construction of K. Hence, the map Φ defined by $\Phi = \Phi_3 \circ \Phi_2 \circ \Phi_1$ is a measurable map that sends $(\mathbb{X}_n, \mathbb{Y}_n)$ to $\mathrm{M}_{r,g,\delta}(\mathbb{X}_n, \mathbb{Y}_n)$.

A.5 Proof of Proposition 3.2

We fix some parameters $a > 0$ and $b \geq 1$. First note that Assumption (2) is always satisfied by definition of r_n. Next, there exists $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$, Assumption (1) is satisfied because $\delta_n \to 0$ and $\omega(\delta_n) \to 0$ as $n \to +\infty$. Moreover, n_0 can be taken the same for all $f \in \bigcup_{\mathbb{P} \in \mathcal{P}(a,b)} \mathcal{F}(\mathbb{P}, \omega)$.

Let $\varepsilon_n := d_{\mathrm{H}}(\mathcal{X}, \mathbb{X}_n)$. Under the (a, b)-standard assumption, it is well known that (see for instance Cuevas and Rodríguez-Casal (2004); Chazal et al. (2015b)):

$$P\left(\varepsilon_n \geq u\right) \leq \min\left\{1, \frac{4^b}{au^b} \mathrm{e}^{-a\left(\frac{u}{2}\right)^b n}\right\}, \forall u > 0. \tag{15}$$

In particular, regarding the complementary of (3) we have:

$$P\left(\varepsilon_n > \frac{\delta_n}{4}\right) \leq \min\left\{1, \frac{2^b}{2\log(n)n}\right\}. \tag{16}$$

Recall that $\mathrm{diam}(\mathcal{X}_{\mathbb{P}}) \leq L$. Let $\bar{C} = \omega(L)$ be a constant that only depends on the parameters of the model. Then, for any $\mathbb{P} \in \mathcal{P}(a, b)$, we have:

$$\sup_{f \in \mathcal{F}(\mathbb{P}, \omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n\right) \leq \bar{C}. \tag{17}$$

For $n \geq n_0$, we have :

$$\sup_{f \in \mathcal{F}(\mathbb{P}, \omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n\right) = \sup_{f \in \mathcal{F}(\mathbb{P}, \omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n\right) \mathbb{I}_{\varepsilon_n > \delta_n/4} + \sup_{f \in \mathcal{F}(\mathbb{P}, \omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n\right) \mathbb{I}_{\varepsilon_n \leq \delta_n/4}$$

[1] This is because singletons are open balls in the discrete topology, and also because of the stability theorem for persistence diagrams Chazal et al. (2016a); Cohen-Steiner et al. (2007).

and thus

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}(\mathbb{P},\omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_\mathbb{P}),\mathrm{M}_n\right)\right] &\leq \bar{C}P\left(\varepsilon_n > \frac{\delta_n}{4}\right) + \sup_{f\in\mathcal{F}(\mathbb{P},\omega)}\left[\frac{r_n}{2} + 2\omega(\delta_n)\right] \\ &\leq \bar{C}\min\left\{1, \frac{2^b}{2\log(n)n}\right\} + \left(\frac{1+4g}{2g}\right)\omega(\delta_n) \qquad (18)\end{aligned}$$

where we have used (17), Theorem 2.7 and the fact that $V_n(\delta_n) \leq \omega(\delta_n)$. For n large enough, the first term in (18) is of the order of δ_n^b, which can be upper bounded by δ_n and thus by $\omega(\delta_n)$ (up to a constant) since $\omega(\delta)/\delta$ is non-increasing. Since $\frac{1+4g}{2g} < \frac{9}{2}$ because $\frac{1}{3} < g < \frac{1}{2}$, we get that the risk is bounded by $\omega(\delta_n)$ for $n \geq n_0$ up to a constant that only depends on the parameters of the model. The same inequality is of course valid for any n by taking a larger constant, because n_0 itself only depends on the parameters of the model.

A.6 Proof of Proposition 3.3

The proof follows closely Section B.2 of Chazal et al. (2013). Let $\mathcal{X}_0 = [0, a^{-1/b}] \subset \mathbb{R}^D$. Obviously, $\mathcal{X}_0$ is a compact submanifold of $\mathbb{R}^D$. Let $\mathcal{U}(\mathcal{X}_0)$ be the uniform measure on $\mathcal{X}_0$. Let $\mathcal{P}_{a,b,\mathcal{X}_0}$ denote the set of (a,b)-standard measures whose support is included in $\mathcal{X}_0$. Let $x_0 = 0 \in \mathcal{X}_0$ and $\{x_n\}_{n\in\mathbb{N}^*} \in \mathcal{X}_0^{\mathbb{N}}$ such that $\|x_n - x_0\| = (an)^{-1/b}$. Now, let

$$f_0 : \left\{ \begin{array}{ll} \mathcal{X}_0 & \to \mathbb{R} \\ x & \mapsto \omega(\|x - x_0\|) \end{array} \right. .$$

By definition, we have $f_0 \in \mathcal{F}(\mathcal{U}(\mathcal{X}_0),\omega)$ because $\mathrm{Dg}(\mathcal{X}_0, f_0) = \{(0, \omega(a^{-1/b}))\}$ since f_0 is increasing by definition of ω, and thus $e_{\min}(\mathcal{X}_0, f_0) = p_{\min}(\mathcal{X}_0, f_0) = +\infty$. Finally, given any measure $\mathbb{P} \in \mathcal{P}_{a,b,\mathcal{X}_0}$, we let $\theta_0(\mathbb{P}) := \mathrm{R}_{f_0|_{\mathcal{X}_\mathbb{P}}}(\mathcal{X}_\mathbb{P})$. Then, we have:

$$\begin{aligned}&\sup_{\mathbb{P}\in\mathcal{P}_{a,b}} \mathbb{E}\left[\sup_{f\in\mathcal{F}(\mathbb{P},\omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_\mathbb{P}), \hat{\mathrm{R}}_n\right)\right] \\ &\qquad \geq \sup_{\mathbb{P}\in\mathcal{P}_{a,b,\mathcal{X}_0}} \mathbb{E}\left[\sup_{f\in\mathcal{F}(\mathbb{P},\omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_\mathbb{P}), \hat{\mathrm{R}}_n\right)\right] \\ &\qquad \geq \sup_{\mathbb{P}\in\mathcal{P}_{a,b,\mathcal{X}_0}} \mathbb{E}\left[d_\Delta\left(\mathrm{R}_{f_0|_{\mathcal{X}_\mathbb{P}}}(\mathcal{X}_\mathbb{P}), \hat{\mathrm{R}}_n\right)\right] = \sup_{\mathbb{P}\in\mathcal{P}_{a,b,\mathcal{X}_0}} \mathbb{E}\left[\rho\left(\theta_0(\mathbb{P}), \hat{\mathrm{R}}_n\right)\right],\end{aligned}$$

where $\rho := d_\Delta$. For any $n \in \mathbb{N}^*$, we let $\mathbb{P}_{0,n} := \delta_{x_0}$ be the Dirac measure on x_0 and $\mathbb{P}_{1,n} := (1 - \frac{1}{n})\mathbb{P}_{0,n} + \frac{1}{n}\mathcal{U}([x_0, x_n])$. As a Dirac measure, $\mathbb{P}_{0,n}$ is obviously in $\mathcal{P}_{a,b,\mathcal{X}_0}$. We now check that $\mathbb{P}_{1,n} \in \mathcal{P}_{a,b,\mathcal{X}_0}$.

- Let us study $\mathbb{P}_{1,n}(B(x_0, r))$.

  Assume $r \leq (an)^{-1/b}$. Then

  $$\mathbb{P}_{1,n}(B(x_0,r)) = 1 - \frac{1}{n} + \frac{1}{n}\frac{r}{(an)^{-1/b}} \geq \left(1 - \frac{1}{n} + \frac{1}{n}\right)\left(\frac{r}{(an)^{-1/b}}\right)^b \geq \left(\frac{1}{2} + \frac{1}{n}\right) anr^b \geq ar^b.$$

  Assume $r > (an)^{-1/b}$. Then

$$\mathbb{P}_{1,n}(B(x_0,r)) = 1 \geq \min\{ar^b\}.$$

- Let us study $\mathbb{P}_{1,n}(B(x_n,r))$. Assume $r \leq (an)^{-1/b}$. Then

$$\mathbb{P}_{1,n}(B(x_n,r)) = \frac{1}{n}\frac{r}{(an)^{-1/b}} \geq \frac{1}{n}\left(\frac{r}{(an)^{-1/b}}\right)^b = ar^b.$$

  Assume $r > (an)^{-1/b}$. Then

$$\mathbb{P}_{1,n}(B(x_n,r)) = 1 \geq \min\{ar^b\}.$$

- Let us study $\mathbb{P}_{1,n}(B(x,r))$, where $x \in (x_0,x_n)$. Assume $r \leq x$. Then

$$\mathbb{P}_{1,n}(B(x,r)) \geq \frac{1}{n}\frac{r}{(ab)^{-1/b}} \geq ar^b \text{ (see previous case).}$$

  Assume $r > x$. Then $\mathbb{P}_{1,n}(B(x,r)) = 1 - \frac{1}{n} + \frac{1}{n}\frac{\left(x + \min\{r,\ (an)^{-1/b} - x\}\right)}{(an)^{-1/b}}$. If $\min\{r,\ (an)^{-1/b} - x\} = r$, then we have

$$\mathbb{P}_{1,n}(B(x,r)) \geq 1 - \frac{1}{n} + \frac{1}{n}\frac{r}{(ab)^{-1/b}} \geq ar^b \text{ (see previous case).}$$

  Otherwise, we have

$$\mathbb{P}_{1,n}(B(x,r)) = 1 \geq \min\{ar^b\}.$$

Thus $\mathbb{P}_{1,n}$ is in $\mathcal{P}_{a,b,\mathcal{X}_0}$ as well. Hence, we apply Le Cam's Lemma (see Section B) to get:

$$\sup_{\mathbb{P}\in\mathcal{P}_{a,b,\mathcal{X}_0}} \mathbb{E}\left[\rho\left(\theta_0(\mathbb{P}), \hat{\mathrm{R}}_n\right)\right] \geq \frac{1}{8}\rho(\theta_0(\mathbb{P}_{0,n}), \theta_0(\mathbb{P}_{1,n}))\left[1 - \mathrm{TV}(\mathbb{P}_{0,n}, \mathbb{P}_{1,n})\right]^{2n}.$$

By definition, we have:

$$\rho(\theta_0(\mathbb{P}_{0,n}), \theta_0(\mathbb{P}_{1,n})) = d_\Delta\left(\mathrm{R}_{f_0|_{\{x_0\}}}(\{x_0\}), \mathrm{R}_{f_0|_{[x_0,x_n]}}(\mathcal{U}[x_0,x_n])\right).$$

Since $\mathrm{Dg}\left(\mathrm{R}_{f_0|_{\{x_0\}}}(\{x_0\})\right) = \{(0,0)\}$ and $\mathrm{Dg}\left(\mathrm{R}_{f_0|_{[x_0,x_n]}}(\mathcal{U}[x_0,x_n])\right) = \{(f(x_0), f(x_n))\}$ because f_0 is increasing by definition of ω, it follows that

$$\rho(\theta_0(\mathbb{P}_{0,n}), \theta_0(\mathbb{P}_{1,n})) = \frac{1}{2}|f(x_n) - f(x_0)| = \frac{1}{2}\omega\left((an)^{-1/b}\right).$$

It remains to compute $\mathrm{TV}(\mathbb{P}_{0,n}, \mathbb{P}_{1,n}) = \left|1 - \left(1 - \frac{1}{n}\right)\right| + \frac{1}{n}(an)^{-1/b} = \frac{1}{n} + o\left(\frac{1}{n}\right)$. The Proposition follows then from the fact that $\left[1 - \mathrm{TV}(\mathbb{P}_{0,n}, \mathbb{P}_{1,n})\right]^{2n} \to \mathrm{e}^{-2}$.

A.7 Proof of Proposition 3.4

Let $\mathbb{P} \in \mathcal{P}_{a,b}$ and ω a modulus of continuity of f. Using the same notation as in the previous section, we have

$$\begin{aligned} P\left(\delta_n \geq u\right) &\leq P\left(d_{\mathrm{H}}(\mathbb{X}_n, \mathcal{X}_{\mathbb{P}}) \geq \frac{u}{2}\right) + P\left(d_{\mathrm{H}}(\mathbb{X}_n^{s_n}, \mathcal{X}_{\mathbb{P}}) \geq \frac{u}{2}\right) \\ &\leq P\left(\varepsilon_n \geq \frac{u}{2}\right) + P\left(\varepsilon_{s_n} \geq \frac{u}{2}\right). \end{aligned} \tag{19}$$

Note that for any $f \in \mathcal{F}(P, \omega)$, according to (4) and (17)

$$d_{\Delta}\left(\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}), \mathrm{M}_n\right) \leq \left[\frac{r}{2} + 2\omega(\delta)\right] \mathbb{I}_{\Omega_n} + \bar{C}\, \mathbb{I}_{\Omega_n^c} \tag{20}$$

where Ω_n is the event defined by

$$\Omega_n = \{4\delta_n \leq \min\{\kappa, \rho\}\} \cap \{2\omega(\delta_n) \leq \underline{e}\} \cap \{4\varepsilon_n \leq \delta_n\}.$$

This gives

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \mathcal{F}(P,\omega)} d_{\Delta}\left(\mathrm{M}_n, \mathrm{R}_f(\mathcal{X})\right)\right] \leq & \underbrace{\int_0^{\bar{C}} P\left(\omega(\delta_n) \geq \frac{2g}{1+4g}\alpha\right) \mathrm{d}\alpha}_{(A)} + \underbrace{\bar{C} P\left(\varepsilon_n \geq \frac{\delta_n}{4}\right)}_{(B)} \\ & + \underbrace{\bar{C} P\left(\omega(\delta_n) \geq \frac{1}{2}\underline{e}\right)}_{(C)} + \underbrace{\bar{C} P\left(\delta_n \geq \min\left\{\frac{\kappa}{4}, \frac{\rho}{4}\right\}\right)}_{(D)}. \end{aligned}$$

Let us bound the four terms (A), (B), (C) and (D).

- **Terms** (C) **and** (D). Both terms can be bounded using (19) then (15). Indeed, since ω is increasing, one has for all $u > 0$:

$$P\left(\omega(\delta_n) \geq u\right) = P\left(\delta_n \geq \omega^{-1}(u)\right). \tag{21}$$

- **Term** (B). Let $t_n = 2\left(\frac{2\log(n)}{an}\right)^{1/b}$ and $A_n = \{\varepsilon_n < t_n\}$. It is known that on the event A_n, one has $\delta_n \geq 4\varepsilon_n$ for n large enough (see for instance Section 6 in Fasy et al. (2014)). Thus, one has:

$$P\left(\varepsilon_n \geq \frac{\delta_n}{4}\right) \leq \underbrace{P\left(\varepsilon_n \geq \frac{\delta_n}{4} \mid A_n\right)}_{=0} P\left(A_n\right) + P\left(A_n^c\right) = P\left(A_n^c\right).$$

  Finally, the probability of A_n^c is bounded with (15):

$$P\left(A_n^c\right) \leq \frac{2^b}{2\log(n)n}.$$

- **Term** (A). This is the dominating term. Using (19) and (21), we have:

$$(A) \leq \int_0^{\bar{C}} P\left(\varepsilon_n \geq \frac{1}{2}\omega^{-1}\left(\frac{2g\alpha}{1+4g}\right)\right) \mathrm{d}\alpha + \int_0^{\bar{C}} P\left(\varepsilon_{s_n} \geq \frac{1}{2}\omega^{-1}\left(\frac{2g\alpha}{1+4g}\right)\right) \mathrm{d}\alpha.$$

We only bound the first integral, but the analysis extends verbatim to the second integral when replacing n by s_n. Let

$$\alpha_n = \frac{1+4g}{2g}\omega\left[\left(\frac{4^b \log(n)}{an}\right)^{1/b}\right].$$

Since $x \mapsto \frac{\omega(x)}{x}$ is non-increasing, it follows that $x \mapsto \frac{\omega^{-1}(x)}{x}$ is non-decreasing, and

$$\omega^{-1}(x) \geq \frac{x}{y}\omega^{-1}(y), \ \forall x \geq y > 0. \tag{22}$$

Using (15), we have the following inequalities:

$$\begin{aligned}
\int_0^{\bar{C}} P\left(\varepsilon_n \geq \frac{1}{2}\omega^{-1}\left(\frac{2g\alpha}{1+4g}\right)\right) \mathrm{d}\alpha &\leq \alpha_n + \frac{8^b}{a}\int_{\alpha_n}^{\bar{C}} \frac{1}{\omega^{-1}\left(\frac{2g\alpha}{1+4g}\right)^b} \exp\left[-\frac{an}{4^b}\omega^{-1}\left(\frac{2g\alpha}{1+4g}\right)^b\right] \mathrm{d}\alpha \\
&\leq \alpha_n + \frac{8^b}{a}\int_{\alpha_n}^{\bar{C}} \frac{\alpha_n^b}{\left[\alpha\omega^{-1}\left(\frac{2g\alpha_n}{1+4g}\right)\right]^b} \exp\left[-\frac{an\alpha^b}{(4\alpha_n)^b}\omega^{-1}\left(\frac{2g\alpha_n}{1+4g}\right)^b\right] \mathrm{d}\alpha \\
&\leq \alpha_n + \alpha_n \frac{2^b 4 n^{1-1/b}}{b a^{1/b} \omega^{-1}\left(\frac{2g\alpha_n}{1+4g}\right)} \int_{u \geq \frac{an}{4^b}\omega^{-1}\left(\frac{2g\alpha_n}{1+4g}\right)^b} u^{1/b-2}\mathrm{e}^{-u}\mathrm{d}u \\
&= \alpha_n + \alpha_n \frac{2^b n}{b\log(n)^{1/b}} \int_{u \geq \log(n)} u^{1/b-2}\mathrm{e}^{-u}\mathrm{d}u \\
&\leq C(a,b)\alpha_n,
\end{aligned}$$

where we used (22) with $x = \frac{2g\alpha}{1+4g}$ and $y = \frac{2g\alpha_n}{1+4g}$ for the second inequality. The constant $C(a,b)$ only depends on a and b.

Hence, since $\frac{1+4g}{2g} < \frac{9}{2}$, there exist constants $K, K' > 0$ that depend only of the geometric parameters of the model such that:

$$(A) \leq K\omega\left(\frac{K'\log(s_n)}{s_n}\right)^{1/b}.$$

Final bound. Since $s_n = n\log(n)^{-(1+\beta)}$, by gathering all four terms, there exist constants $C, C' > 0$ such that:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}(\mathbb{P},\omega)} d_\Delta\left(\mathrm{R}_f(\mathcal{X}_\mathbb{P}), \mathrm{M}_n\right)\right] \leq C\omega\left(\frac{C'\log(n)^{2+\beta}}{n}\right)^{1/b}.$$

A.8 Proof of Proposition 4.1

We have the following bound by using (20) in the proof of Proposition 3.4:

$$\begin{aligned}
&P\left(d_{\Delta}(,\mathrm{R}_f(\mathcal{X}_{\mathbb{P}}),\mathrm{M}_n)\geq\eta\right)\\
&\leq P\left(\omega(\delta_n)\geq\frac{2g}{1+4g}\eta\right)+P\left(\varepsilon_n\geq\frac{\delta_n}{4}\right)\\
&\qquad+P\left(\omega(\delta_n)\geq\frac{1}{2}\underline{e}\right)+P\left(\delta_n\geq\min\left\{\frac{\kappa}{4},\frac{\rho}{4}\right\}\right)\\
&\leq P\left(\varepsilon_n\geq\frac{1}{2}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right)+P\left(\varepsilon_{s_n}\geq\frac{1}{2}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right)+o\left(\frac{1}{n\log(n)}\right).
\end{aligned}$$

Following the lines of Section 6 in Fasy et al. (2014), subsampling approximations give

$$P\left(\varepsilon_n\geq\frac{1}{2}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right)\leq L_n\left(\frac{1}{4}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right)+o\left(\frac{s_n}{n}\right)^{1/4},$$

and

$$P\left(\varepsilon_{s_n}\geq\frac{1}{2}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right)\leq F_n\left(\frac{1}{4}\omega^{-1}\left(\frac{2g}{1+4g}\eta\right)\right)+o\left(\frac{s_n^2}{s_n}\right)^{1/4}.$$

The result follows by taking $s_n=n\log(n)^{-(1+\beta)}$.

B Le Cam's Lemma

The version of Le Cam's Lemma given below is from Yu (1997) (see also Genovese et al., 2012b). Recall that the total variation distance between two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ on a measured space $(\mathcal{X},\mathcal{B})$ is defined by

$$\mathrm{TV}(\mathbb{P}_0,\mathbb{P}_1)=\sup_{B\in\mathcal{B}}|\mathbb{P}_0(B)-\mathbb{P}_1(B)|.$$

Moreover, if $\mathbb{P}_0$ and $\mathbb{P}_1$ have densities p_0 and p_1 for the same measure λ on $\mathcal{X}$, then

$$\mathrm{TV}(\mathbb{P}_0,\mathbb{P}_1)=\frac{1}{2}\ell_1(p_0,p_1):=\int_{\mathcal{X}}|p_0-p_1|d\lambda.$$

Lemma B.1. *Let $\mathcal{P}$ be a set of distributions. For $\mathbb{P}\in\mathcal{P}$, let $\theta(\mathbb{P})$ take values in a pseudometric space $(\mathbb{X},\rho)$. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ in $\mathcal{P}$ be any pair of distributions. Let $X_1,\ldots,X_n$ be drawn i.i.d. from some $\mathbb{P}\in\mathcal{P}$. Let $\hat{\theta}=\hat{\theta}(X_1,\ldots,X_n)$ be any estimator of $\theta(\mathbb{P})$, then*

$$\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}_{\mathbb{P}^n}\left[\rho(\theta,\hat{\theta})\right]\geq\frac{1}{8}\rho\left(\theta(\mathbb{P}_0),\theta(\mathbb{P}_1)\right)\left[1-\mathrm{TV}(\mathbb{P}_0,\mathbb{P}_1)\right]^{2n}.$$

C Extended Persistence

Let f be a real-valued function on a topological space X. The family $\{X^{(-\infty,\alpha]}\}_{\alpha\in\mathbb{R}}$ of sublevel sets of f defines a *filtration*, that is, it is nested w.r.t. inclusion: $X^{(-\infty,\alpha]}\subseteq X^{(-\infty,\beta]}$ for all $\alpha\leq\beta\in\mathbb{R}$. The family $\{X^{[\alpha,+\infty)}\}_{\alpha\in\mathbb{R}}$ of superlevel sets of f is also nested but in the opposite direction: $X^{[\alpha,+\infty)}\supseteq X^{[\beta,+\infty)}$ for all $\alpha\leq\beta\in\mathbb{R}$. We can turn it into a filtration by reversing the

real line. Specifically, let $\mathbb{R}^{\text{op}} = \{\tilde{x} \mid x \in \mathbb{R}\}$, ordered by $\tilde{x} \leq \tilde{y} \Leftrightarrow x \geq y$. We index the family of superlevel sets by $\mathbb{R}^{\text{op}}$, so now we have a filtration: $\{X^{[\tilde{\alpha},+\infty)}\}_{\tilde{\alpha}\in\mathbb{R}^{\text{op}}}$, with $X^{[\tilde{\alpha},+\infty)} \subseteq X^{[\tilde{\beta},+\infty)}$ for all $\tilde{\alpha} \leq \tilde{\beta} \in \mathbb{R}^{\text{op}}$.

Extended persistence connects the two filtrations at infinity as follows. Replace each superlevel set $X^{[\tilde{\alpha},+\infty)}$ by the pair of spaces $(X, X^{[\tilde{\alpha},+\infty)})$ in the second filtration. This maintains the filtration property since we have $(X, X^{[\tilde{\alpha},+\infty)}) \subseteq (X, X^{[\tilde{\beta},+\infty)})$ for all $\tilde{\alpha} \leq \tilde{\beta} \in \mathbb{R}^{\text{op}}$. Then, let $\mathbb{R}_{\text{Ext}} = \mathbb{R} \cup \{+\infty\} \cup \mathbb{R}^{\text{op}}$, where the order is completed by $\alpha < +\infty < \tilde{\beta}$ for all $\alpha \in \mathbb{R}$ and $\tilde{\beta} \in \mathbb{R}^{\text{op}}$. This poset is isomorphic to $(\mathbb{R}, \leq)$. Finally, define the *extended filtration* of f over $\mathbb{R}_{\text{Ext}}$ by:

$$\begin{aligned} F_\alpha &= X^{(-\infty,\alpha]} && \text{for } \alpha \in \mathbb{R} \\ F_{+\infty} &= X \equiv (X, \emptyset) \\ F_{\tilde{\alpha}} &= (X, X^{[\tilde{\alpha},+\infty)}) && \text{for } \tilde{\alpha} \in \mathbb{R}^{\text{op}}, \end{aligned}$$

where we have identified the space X with the pair of spaces $(X, \emptyset)$. This is a well-defined filtration since we have $X^{(-\infty,\alpha]} \subseteq X \equiv (X, \emptyset) \subseteq (X, X^{[\tilde{\beta},+\infty)})$ for all $\alpha \in \mathbb{R}$ and $\tilde{\beta} \in \mathbb{R}^{\text{op}}$. The subfamily $\{F_\alpha\}_{\alpha\in\mathbb{R}}$ is called the *ordinary* part of the filtration, and the subfamily $\{F_{\tilde{\alpha}}\}_{\tilde{\alpha}\in\mathbb{R}^{\text{op}}}$ is called the *relative* part. See Figure 9 for an illustration.

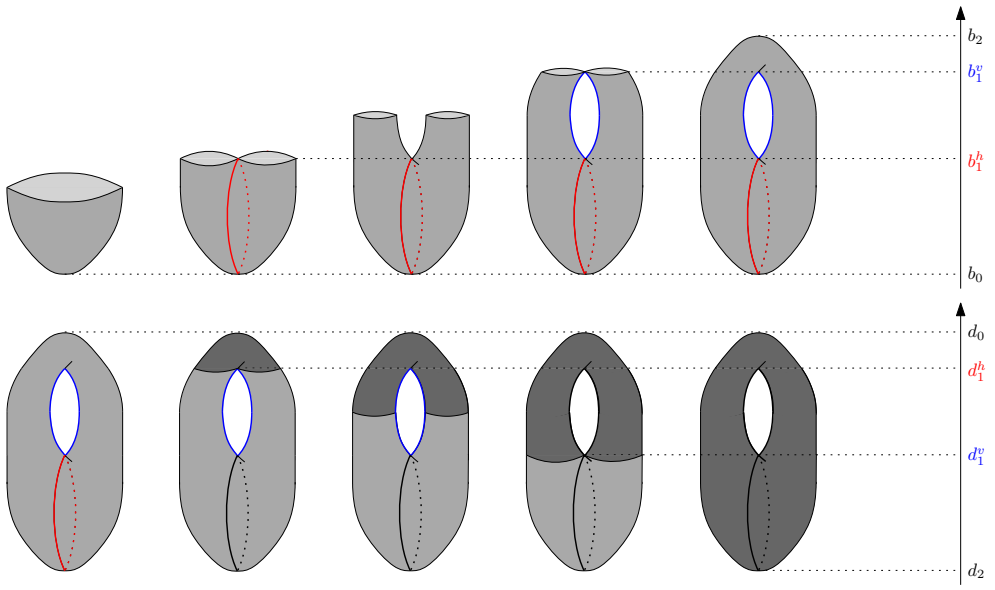

Figure 9: The extended filtration of the height function on a torus. The upper row displays the ordinary part of the filtration while the lower row displays the relative part. The red and blue cycles both correspond to extended points in dimension 1. The point corresponding to the red cycle is located above the diagonal ($d_1^h > b_1^h$), while the point corresponding to the blue cycle is located below the diagonal ($d_1^v > b_1^v$).

Applying the homology functor H_* to this filtration gives the so-called *extended persistence*

module $\mathbb{V}$ of f:

$$\begin{aligned} V_\alpha &= H_*(F_\alpha) = H_*(X^{(-\infty,\alpha]}) && \text{for } \alpha \in \mathbb{R} \\ V_{+\infty} &= H_*(F_{+\infty}) = H_*(X) \cong H_*(X, \emptyset) \\ V_{\tilde{\alpha}} &= H_*(F_{\tilde{\alpha}}) = H_*(X, X^{[\tilde{\alpha},+\infty)}) && \text{for } \tilde{\alpha} \in \mathbb{R}^{\mathrm{op}}, \end{aligned}$$

and where the linear maps between the spaces are induced by the inclusions in the extended filtration.

For Morse-type functions, the extended persistence module can be decomposed as a finite direct sum of half-open *interval modules*—see e.g. Chazal et al. (2016a):

$$\mathbb{V} \simeq \bigoplus_{k=1}^{n} \mathbb{I}[b_k, d_k),$$

where each summand $\mathbb{I}[b_k, d_k)$ is made of copies of the field of coefficients at each index $\alpha \in [b_k, d_k)$, and of copies of the zero space elsewhere, the maps between copies of the field being identities. Each summand represents the lifespan of a *homological feature* (cc, hole, void, etc.) within the filtration. More precisely, the *birth time* b_k and *death time* d_k of the feature are given by the endpoints of the interval. Then, a convenient way to represent the structure of the module is to plot each interval in the decomposition as a point in the extended plane, whose coordinates are given by the endpoints. Such a plot is called the *extended persistence diagram* of f, denoted $\mathrm{Dg}(f)$. The distinction between ordinary and relative parts of the filtration allows to classify the points in $\mathrm{Dg}(f)$ in the following way:

- points whose coordinates both belong to $\mathbb{R}$ are called *ordinary* points; they correspond to homological features being born and then dying in the ordinary part of the filtration;
- points whose coordinates both belong to $\mathbb{R}^{\mathrm{op}}$ are called *relative* points; they correspond to homological features being born and then dying in the relative part of the filtration;
- points whose abscissa belongs to $\mathbb{R}$ and whose ordinate belongs to $\mathbb{R}^{\mathrm{op}}$ are called *extended* points; they correspond to homological features being born in the ordinary part and then dying in the relative part of the filtration.

Note that ordinary points lie strictly above the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$ and relative points lie strictly below Δ, while extended points can be located anywhere, including on Δ, e.g. cc that lie inside a single critical level. It is common to decompose $\mathrm{Dg}(f)$ according to this classification:

$$\mathrm{Dg}(f) = \mathrm{Ord}(f) \sqcup \mathrm{Rel}(f) \sqcup \mathrm{Ext}^+(f) \sqcup \mathrm{Ext}^-(f),$$

where by convention $\mathrm{Ext}^+(f)$ includes the extended points located on the diagonal Δ.

References

Biau, G. and Mas, A. (2012). PCA-Kernel estimation. *Statistics and Risk Modeling with Applications in Finance and Insurance*, 29(1):19–46.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294.

Buchet, M., Chazal, F., Oudot, S., and Sheehy, D. (2015). Efficient and Robust Persistent Homology for Measures. In *Proceedings of the 26th Symposium on Discrete Algorithms*, pages 168–180.

Carrière, M. and Oudot, S. (2015). Structure and Stability of the 1-Dimensional Mapper. *CoRR*, abs/1511.05823.

Carrière, M. and Oudot, S. (2016). Structure and Stability of the 1-Dimensional Mapper. In *Proceedings of the 32nd Symposium on Computational Geometry*, volume 51, pages 25:1–25:16.

Carrière, M. and Oudot, S. (2017). Local Equivalence and Induced Metrics for Reeb Graphs. In *Proceedings of the 33rd Symposium on Computational Geometry*.

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011). Geometric Inference for Probability Measures. *Foundations of Computational Mathematics*, 11(6):733–751.

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2016a). *The Structure and Stability of Persistence Modules*. Springer.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2014). Robust topological inference: distance to a measure and kernel distance. *CoRR*, abs/1412.7197. Accepted for publication in Journal of Machine Learning Research.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2015a). Subsampling Methods for Persistent Homology. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2143–2151.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2013). Optimal rates of convergence for persistence diagrams in Topological Data Analysis. *CoRR*, abs/1305.6239.

Chazal, F., Glisse, M., Labruère, C., and Michel, B. (2015b). Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16:3603–3635.

Chazal, F., Massart, P., and Michel, B. (2016b). Rates of convergence for robust geometric inference. *Electronic Journal of Statistics*, 10(2):2243–2286.

Chen, X., Golovinskiy, A., and Funkhouser, T. (2009). A Benchmark for 3D Mesh Segmentation. *ACM Transactions on Graphics*, 28(3):1–12.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of Persistence Diagrams. *Discrete and Computational Geometry*, 37(1):103–120.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2009). Extending persistence using Poincaré and Lefschetz duality. *Foundation of Computational Mathematics*, 9(1):79–103.

Cuevas, A. (2009). Set estimation: another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa*, 25(2):71–85.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability*, pages 340–354.

DeVore, R. and Lorentz, G. (1993). *Constructive approximation*, volume 303. Springer Science & Business Media.

Dey, T. and Wang, Y. (2013). Reeb Graphs: Approximation and Persistence. *Discrete and Computational Geometry*, 49(1):46–73.

Fasy, B., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). Confidence Sets for Persistence Diagrams. *The Annals of Statistics*, 42(6):2301–2339.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer series in statistics Springer, Berlin.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012a). Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40:941–963.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012b). Minimax Manifold Estimation. *Journal of Machine Learning Research*, 13:1263–1291.

Hinks, T., Zhou, X., Staples, K., Dimitrov, B., Manta, A., Petrossian, T., Lum, P., Smith, C., Ward, J., Howarth, P., Walls, A., Gadola, S., and Djukanovic, R. (2015). Innate and adaptive t cells in asthmatic patients: Relationship to severity and disease mechanisms. *Journal of Allergy and Clinical Immunology*, 136(2):323–333.

Lum, P., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3.

Morozov, D. (2008). *Homological Illusions of Persistence and Stability*. Ph.D. dissertation, Department of Computer Science, Duke University.

Nene, S., Nayar, S., and Murase, H. (1996). Columbia Object Image Library (COIL-100). Technical Report CUCS-006-96.

Nielson, J., Paquette, J., Liu, A., Guandique, C., Tovar, A., Inoue, T., Irvine, K.-A., Gensel, J., Kloke, J., Petrossian, T., Lum, P., Carlsson, G., Manley, G., Young, W., Beattie, M., Bresnahan, J., and Ferguson, A. (2015). Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, 6.

Reaven, G. and Miller, R. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16(1):17–24.

Rucco, M., Merelli, E., Herman, D., Ramanan, D., Petrossian, T., Falsetti, L., Nitti, C., and Salvi, A. (2015). Using topological data analysis for diagnosis pulmonary embolism. *Journal of Theoretical and Applied Computer Science*, 9(1):41–55.

Sarikonda, G., Pettus, J., Phatak, S., Sachithanantham, S., Miller, J., Wesley, J., Cadag, E., Chae, J., Ganesan, L., Mallios, R., Edelman, S., Peters, B., and von Herrath, M. (2014). Cd8 t-cell reactivity to islet antigens is unique to type 1 while cd4 t-cell reactivity exists in both type 1 and type 2 diabetes. *Journal of Autoimmunity*, 50:77–82.

Shawe-Taylor, J., Williams, C., Cristianini, N., and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522.

Singh, G., Mémoli, F., and Carlsson, G. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Symposium on Point Based Graphics*, pages 91–100.

The GUDHI Project (2015). *GUDHI User and Reference Manual*. GUDHI Editorial Board.

Yao, Y., Sun, J., Huang, X., Bowman, G., Singh, G., Lesnick, M., Guibas, L., Pande, V., and Carlsson, G. (2009). Topological methods for exploring low-density states in biomolecular folding pathways. *Journal of Chemical Physics*, 130(14).

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.