Journal of Machine Learning Research 24 (2023) 1-35 Submitted 12/21; Revised 11/22; Published 2/23

Outlier-Robust Subsampling Techniques for Persistent Homology

Bernadette J. Stolz[1,2] BERNADETTE.STOLZ-PRETZER@EPFL.CH
[1] *École Polytechnique Fédérale de Lausanne*
Station 8, 1015 Lausanne, Switzerland
[2] *Mathematical Institute*
University of Oxford, Woodstock Rd
Oxford OX2 6GG, United Kingdom

Editor: Sayan Mukherjee

Abstract

In recent years, persistent homology (PH) has been successfully applied to real-world data in many different settings. Despite significant computational advances, PH algorithms do not yet scale to large datasets preventing interesting applications. One approach to address computational issues posed by PH is to select a set of landmarks by subsampling from the data. Currently, these landmark points are chosen either at random or using the maxmin algorithm. Neither is ideal as random selection tends to favour dense areas of the data while the maxmin algorithm is very sensitive to noise. Here, we propose a novel approach to select landmarks specifically for PH that preserves coarse topological information of the original dataset. Our method is motivated by the Mayer-Vietoris sequence and requires only local PH calculations thus enabling efficient computation. We test our landmarks on artificial data sets which contain different levels of noise and compare them to standard landmark selection techniques. We demonstrate that our landmark selection outperforms standard methods as well as a subsampling technique based on an outlier-robust version of the k-means algorithm for low sampling densities in noisy data with respect to robustness to outliers.

Keywords: landmarks, persistent homology, subsampling, outliers, noise

1. Introduction

One of the practical challenges of applying persistent homology (PH) is that it is computationally difficult for large data sets. Building a filtration on a data set with N points results in spaces of the size $\mathcal{O}(2^N)$, although in practice this can be reduced to $\mathcal{O}(N^{\hat{n}+1})$ by posing a limit $\hat{n}$ on the dimension of the topological features considered (Otter et al., 2017). Following the construction of a filtration, the algorithm for computing PH in the worst case has a complexity of $\mathcal{O}(\mathtt{k}^3)$, where $\mathtt{k}$ is the number of points, edges, triangles, and higher-dimensional simplices constructed on the data points in the filtration. In practice, the complexity is often linear (Otter et al., 2017; Edelsbrunner et al., 2002). Methods exist to approximate specific filtrations or to reduce the sizes of the vector spaces associated to the data by a filtration, for an overview of such methods that have been implemented in software packages, see Otter et al. (2017). Despite such improvements, computation is still very challenging on large data sets and can pose a hard limit on the filtrations that can be applied to a particular data set. In such cases it can become necessary to preprocess data before applying PH. For point cloud data one can, for example, use subsampling techniques to identify so-called *landmarks* of the data set and then define a filtration on the landmarks (de Silva and Carlsson, 2004). A filtration on well-chosen landmarks retains topologically important global information about the full data set and one can even choose to include information from non-landmark points when constructing its simplices. An example for such a filtration is the *lazy witness filtration* which was first introduced by de Silva and Carlsson (2004) and has been used to study noisy artificial data sets (Kovacev-Nikolic, 2012), primary visual cortex cell populations (Singh et al., 2008), and cancer gene expression data (Lockwood and Krishnamoorthy, 2015). Roughly, the lazy witness filtration consists of the following steps:[1]

1. Selection of a (small) subset[2] of *landmark points* L from the data set D.

2. Construction of a *lazy witness filtration* on the landmarks L where the landmarks are vertices and data points from the full data set D can serve as *witnesses* for higher-order interactions, i.e. simplices, on the set of landmarks in the filtration.

Even though the choice of landmarks from the data inevitably has a large influence on the results that can be obtained, currently there are only two standard approaches to select landmarks: uniform random selection and selection via the *maxmin* algorithm. Neither is ideal and in particular the maxmin algorithm tends to include outliers (de Silva and Carlsson, 2004; Adams and Tausz, 2015), which poses problems since large real-world data sets often include noise. In the biological application of the lazy witness filtration by Lockwood and Krishnamoorthy (2015), for example, the maxmin algorithm leads to the discovery of loops in the data set which we could not reproduce using uniform random landmark selection or when discarding a small proportion of the initially chosen maxmin landmarks from the data set and choosing a new set of landmarks with the maxmin algorithm. We also note that the results of the lazy witness filtration are difficult to interpret. In addition to the mentioned disadvantages of the maxmin algorithm, neither of the proposed landmark selection methods were designed specifically for PH. While there are other methods that address subsampling for PH (Cohen-Steiner et al., 2007; Niyogi et al., 2008; Dufresne et al., 2019) these do not explicitly consider noisy data. Because existing approaches are not ideal, it is desirable to develop new methods that lead to a reduction of large and noisy data while preserving topological properties. Ideally, the reduced data set can be used as input for PH directly without additional preprocessing steps.

While one of the appeals of PH is its ability to study multi-scale data sets globally, local PH around a data point can also produce useful insights, see for example Bendich et al. (2012); Ahmed et al. (2014); Fasy and Wang (2016); Stolz et al. (2020); Wheeler et al. (2021). Here, we present a novel landmark selection technique that allows us to obtain outlier-robust PH landmarks from noisy point cloud data. Our method is motivated by the Mayer-Vietoris sequence and relies on computing the Vietoris-Rips filtration (Carlsson, 2009; Ghrist, 2008) locally, in a small neighbourhood around each data point, where the

[1] For the full definition see de Silva and Carlsson (2004).

[2] Although there is no systematic lower bound for the number of landmark points, (de Silva and Carlsson, 2004) suggest using $> 5\%$ of the data points as landmarks.

aforementioned computational problems of PH vanish. Our algorithm then uses PH output to define a score for each data point which allows us to identify suitable candidates for landmarks. We investigate two different flavours of our approach, one using high scores to identify candidate landmarks and one using low scores. We apply our approach to very simple artificial data sets that consist of signal points sampled from an object with a topologically interesting structure, such as a sphere, a torus, or a Klein bottle, as well as noise points. A consequence of the inclusion of noise is that a large data set cannot be reduced by applying subsampling techniques developed to infer the (persistent) homology of data (Cohen-Steiner et al., 2007; Niyogi et al., 2008; Dufresne et al., 2019). The (global) PH of landmarks selected in our proposed method is close to the PH of signal points in the original data set. In comparison to existing landmark selection procedures, we demonstrate that our landmarks based on local PH perform very well on our data sets, in particular for low sampling densities, with respect to robustness to outliers. We further investigate the performance of our landmarks in comparison with an outlier-robust version of the k-means algorithm (Chawla and Gionis, 2013) which we modify for landmark selection and again find that they are superior for low sampling densities. Our PH landmarks present a valuable addition to the two existing techniques for landmark selection for PH and we believe that they may also be of interest in a broader data science context.

Local homology has previously been used to infer the stratification of data (Bendich et al., 2012; Mileyko, 2021; Nanda, 2020; Robinson et al., 2018; Stolz et al., 2020) as well as road network analysis (Ahmed et al., 2014; Fasy and Wang, 2016) and, more recently, the analysis of topological complexity of data as it passes through different layers of neural networks (Wheeler et al., 2021). A Mayer-Vietoris sequence similar to the one that motivates our landmark selection approach has been used to show Wasserstein stability bounds of the Vietoris-Rips filtration on finite point clouds (Skraba and Turner, 2020) and local computations motivated by the Mayer-Vietoris sequence have also been applied with the goal of parallelising PH computations (Casas, 2019). Although the idea behind our notion of local persistent homology is similar to some of those used previously (Ahmed et al., 2014; Bendich et al., 2012; Fasy and Wang, 2016; Wheeler et al., 2021), in comparison to these approaches, our definition does not use relative homology and can be readily computed by considering the Vietoris-Rips complex on points in a small local neighbourhood.

Our paper is structured as follows: in Section 2, we review existing standard techniques for landmark selection. We then mathematically motivate and introduce our novel landmark selection technique, which we refer to as PH landmarks, and also present an outlier-robust version of the k-means algorithm (Chawla and Gionis, 2013) modified for landmark selection, $k--$ landmarks, in Section 3. In Section 4 we introduce our data sets. We compare our PH landmarks to existing techniques as well as $k--$ landmarks in Section 5 followed by a discussion in Section 6.

2. Existing Landmark Selection Methods

The two standard methods for selecting landmarks $L \subset D$ in a data set $D = \{y_1, \ldots, y_N\}$ are random landmark selection and the maxmin algorithm. Both are implemented as standard procedures for use in combination with the lazy witness filtration in the PH software package javaPlex (Tausz et al., 2014).

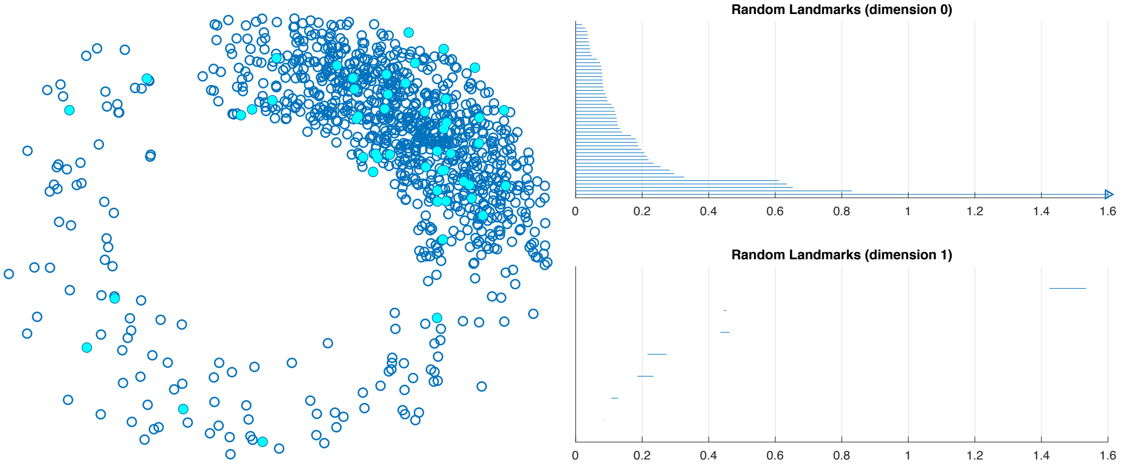


Figure 1: Example of a point cloud and landmarks selected at random (landmarks are shown in cyan). We observe that the dimension 1 barcode based on a Vietoris-Rips filtration on the selected landmarks does not capture the persistent homology of the point cloud correctly.

2.1 Random Landmark Selection

The simplest way to choose landmarks $L = \{l_1, l_2, \ldots, l_m\}$ from a point cloud D is to select m points from D uniformly at random. For data sets whose points are evenly distributed, random selection achieves good coverage at a small computational cost. However, as soon as there are large differences in the density of the data, random selection will favour points from more dense regions, which can result in landmarks that do not represent the point cloud well. In extreme cases, the landmarks do not carry any topological similarity to the original point cloud. We show such an example in Figure 1.

2.2 The Maxmin Algorithm

The sequential maxmin algorithm chooses the first landmark $l_1 \in D$ randomly. Inductively, for $i \geq 2$ and a landmark set $L_{i-1} = \{l_1, l_2, \ldots, l_{i-1}\}$, the algorithm selects the next landmark $l_i \in D \backslash L_{i-1}$ such that for a chosen metric $d : D \times D \to \mathbb{R}$ with $d(y, L) = \min_{l \in L} d(y, l)$ the function mapping

$$y \mapsto d(y, L_{i-1}),$$

is maximised for $y \in D$. We show pseudocode for the procedure in Algorithm 3 in Appendix A.1. The method has been used successfully for image data by Adams and Carlsson (2009); Carlsson et al. (2008). Landmarks chosen in this way cover the data set well, are evenly spaced, and represent the underlying topological features better than landmarks selected at random. The algorithm does, however, tend to include outliers (de Silva and Carlsson, 2004; Adams and Tausz, 2015). We show an example of a point cloud where the selected landmarks do not represent the underlying topology of the data correctly in Figure 2.

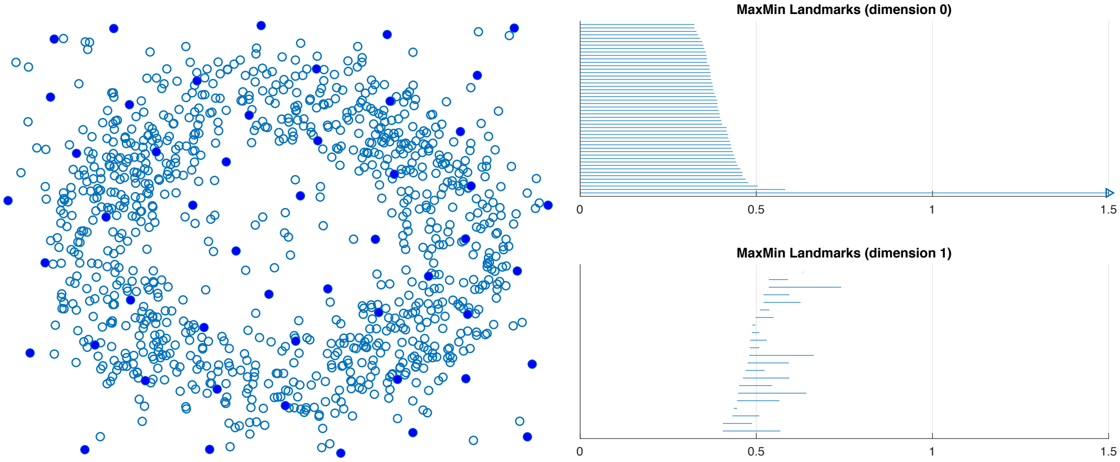


Figure 2: Example of a point cloud and landmarks selected by the maxmin algorithm (landmarks are shown in dark blue). We observe that the dimension 1 barcode based on a Vietoris-Rips filtration on the selected landmarks does not capture the PH of the point cloud correctly.

According to de Silva and Carlsson (2004), the maxmin algorithm is best suited to produce the qualities desired from landmarks. They do not recommend the use of clustering algorithms as an alternative due to the high computational cost and potential to accentuate accidental features.

2.3 Dense Core Subsets

The tutorial accompanying the PH software package javaPlex (Adams and Tausz, 2015) mentions that using so-called dense core subsets before applying the maxmin algorithm can help overcome the selection of outliers as landmarks. This approach, however, is not considered as one of the standard approaches for landmark selection and we did not find examples where it was used in practice for this purpose. De Silva and Carlsson (2004) and Carlsson et al. (2008) apply dense core subsets to identify dense regions in their data sets. The authors subsequently use the maxmin landmarks to study the topology of these dense regions.

Dense core subsets are based on assigning density values to every point: for an integer K, the density value assigned to a point $y \in D$ is $\frac{1}{\rho_K(y)}$, where $\rho_K(y)$ is the distance to the K-th nearest neighbour of y. Large values of K provide a measure of the global density around the point in the data set, while smaller values of K give a more local perspective. Using the density values, one can select the m densest points in the data set as a dense core subset. Given a data set, it is not clear what values of K to use. Different values for K and m can produce markedly different subsets (de Silva and Carlsson, 2004).

For our comparisons in Subsection 5.3, instead of selecting a dense subset and then performing the maxmin algorithm to select landmarks, as proposed by Adams and Tausz (2015), we choose the m densest points in the data as landmarks. This enables us to determine how the information on which our own landmark selection technique (PH landmarks) is based differs from that given by the K-th nearest neighbour of a data point.

3. Proposed Landmark Selection Methods

Currently existing methods for landmark selection that enable subsequent PH analysis on large and noisy point cloud data are not ideal and have, in particular, not been designed with PH in mind. De Silva and Carlsson (2004) state the most pertinent qualities for a landmark set to be good coverage of the data set and even spacing of the landmarks. While these are certainly important properties for many data sets, we find that in the context of PH, the aim of a landmark selection method should be to represent the underlying topology of the data set. Such landmarks could then be used either in combination with the lazy witness complex or simply as a subsample of the data set to which PH can be applied. Since outliers can artificially introduce topological features such as loops, we in particular require outlier-robust landmark selection techniques. Based on these observations we formulate the following goals for landmark selection methods intended for PH:

1. Good representation of the underlying topology of the data set, even at low sampling densities with small variance of the results between different landmark realisations.

2. Robustness to outliers, ideally including a measure for how much we consider a specific point to be an outlier.

We now introduce novel landmark selection methods to achieve these goals: Persistent homology landmarks (PH landmarks) and $k--$ landmarks. We design PH landmarks specifically for the application of PH, while $k--$ landmarks are based on a variant of the k-means algorithm, whose outlier-robust properties make it a promising candidate to overcome the downsides of both the random and the maxmin landmark selection.

3.1 Mathematical Motivation

The use of our notion of local PH is inspired by the Mayer-Vietoris sequence, which can enable computation of the homology of a space X by considering subspaces, whose homology is easier to compute. The Mayer-Vietoris sequence for a topological space X is a special type of long exact sequence. By long exact sequence we mean a sequence of abelian groups A_i and group homomorphisms Φ_i, $i \in \mathbb{Z}$, of the form

$$\cdots \to \mathcal{A}_i \xrightarrow{\Phi_i} \mathcal{A}_{i+1} \xrightarrow{\Phi_{i+1}} \mathcal{A}_{i+2} \xrightarrow{\Phi_{i+2}} \ldots$$

such that $\operatorname{im} \Phi_i = \ker \Phi_{i+1}$. For more definitions and background, see for example Munkres (1984).

Theorem 1 (Mayer-Vietoris sequence) *Let X be a simplicial complex with subcomplexes $A, B \subset X$ such that $X = A \cup B$. Then there exists an exact sequence*

$$\cdots \to H_n(A \cap B) \xrightarrow{\Phi_*} H_n(A) \oplus H_n(B) \xrightarrow{\Psi_*} H_n(X) \xrightarrow{\partial_*} H_{n-1}(A \cap B) \to \ldots$$
$$\to H_0(X) \to 0,$$

where $H_n(\cdot)$ denotes the n-th homology group and Φ_*, Ψ_*, and ∂_* are homomorphisms. The sequence is called the Mayer-Vietoris sequence*.*

For a proof see, for example, Munkres (1984), page 142. We can use the Mayer-Vietoris sequence to connect the homology of a simplicial complex X to the local homology around a vertex $\hat{x} \in X$. We do this via two simplicial subcomplexes that can be defined around the vertex $\hat{x} \in X$: the link of the vertex $\hat{x} \in X$ and the closed star of the vertex $\hat{x} \in X$.

Definition 2 (Closed star of a vertex) *Let X be a simplicial complex and $\hat{x} \in X$ a vertex. Then the closed star of $\hat{x}$ in X, denoted by $\overline{\text{St}}\ \hat{x}$, is the subcomplex of X that contains all the simplices which have $\hat{x}$ as one of their vertices.*

Definition 3 (Link of a vertex) *Let X be a simplicial complex and $\hat{x} \in X$ a vertex. Then the link* $\text{Lk}\ \hat{x}$ *is the union of all simplices of X lying in $\overline{\text{St}}\ \hat{x}$ that are disjoint from $\hat{x}$.*

We show a simplicial complex, the closed star of a vertex and the link of a vertex in Figure 3. Denoting the simplicial complex of all simplices in X that are disjoint from $\hat{x}$ as $X \setminus \hat{x}$, we make the following observations for $\hat{x} \in X$:

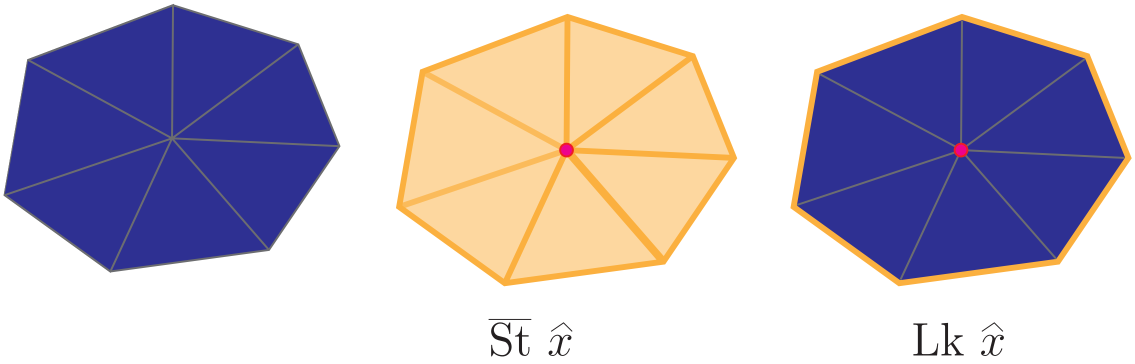


Figure 3: Examples of a simplicial complex, the closed star of a vertex $\hat{x}$, and the link of a vertex $\hat{x}$. We show the vertex $\hat{x}$ in red and highlight the closed star and the link in yellow.

Remark 4

1. $X = (X \setminus \hat{x}) \cup \overline{\text{St}}\ \hat{x}$*.*

2. $\text{Lk}\ \hat{x} = (X \setminus \hat{x}) \cap \overline{\text{St}}\ \hat{x}$*.*

Based on these definitions and observations we can now consider the following Mayer-Vietoris sequence:

$$\begin{aligned} \cdots \to H_n(\text{Lk}\ \hat{x}) \xrightarrow{\Phi} H_n(X \setminus \hat{x}) \oplus H_n(\overline{\text{St}}\ \hat{x}) \xrightarrow{\Psi} H_n(X) \xrightarrow{\partial} H_{n-1}(\text{Lk}\ \hat{x}) \to \dots \\ \to H_0(X) \to 0. \end{aligned} \tag{1}$$

Now, $\overline{\mathrm{St}}\ \hat{x}$ is contractible: every simplex that contains $\hat{x}$ is contractible and the intersection of simplices in $\overline{\mathrm{St}}\ \hat{x}$ is either empty or a simplex that contains $\hat{x}$ and is hence also contractible. Thus $H_n(\overline{\mathrm{St}}\ \hat{x}) = 0$ for $n > 0$. We observe that if we can ensure that $H_n(\mathrm{Lk}\ \hat{x}) = H_{n-1}(\mathrm{Lk}\ \hat{x}) = 0$, for $n > 0$ we obtain

$$0 \to H_n(X \setminus \hat{x}) \xrightarrow{\Psi} H_n(X) \to 0, \tag{2}$$

which gives us an isomorphism Ψ between $H_n(X \setminus \hat{x})$ and $H_n(X)$.

The Mayer-Vietoris sequence given by Equation 1 connects the homology of the large simplicial complexes X and $X \setminus \hat{x}$, which are both global and expensive to compute, to the homology of the link $\mathrm{Lk}\ \hat{x}$, which a purely local computation and therefore easy to perform.

In practice, we are however working with point cloud data D. We therefore consider the PH of the point cloud D, rather than the homology, and apply the Mayer-Vietoris sequence 1 to the simplicial complexes that we obtain from the data by constructing a filtration. We consider a simplicial complex X in this filtration. A data point $y \in D$ is now a vertex $\hat{y}$ in X. Instead of $H_n(\mathrm{Lk}\ \hat{y}) = 0$, we need to quantify the failure of the isomorphism $PH_n(D \setminus y) \xrightarrow{\Psi} PH_n(D)$ via a score based on the $PH_n(\mathrm{Lk}\ y)$, which we will define in Section 3.2. To make computation easier, instead of looking just at the link of $\hat{y}$ in the simplicial complex X, we extend the link to a δ-neighbourhood of $\hat{y}$ in X, which we define to be the collection of simplices whose vertices are within a distance of at most δ from y in D:

Definition 5 (δ-link of a data point y) *Let X be a simplicial complex in a metric space constructed from a data set D, $\hat{y} \in X$ a vertex, $\delta > 0$ a distance and $\delta(\hat{y})$ a δ-neighbourhood of $\hat{y}$ in X. Then the δ-link $\mathrm{Lk}^{\delta}\ y$ is the union of all simplices in X that are disjoint from $\hat{y}$ and contained in $\delta(\hat{y})$.*

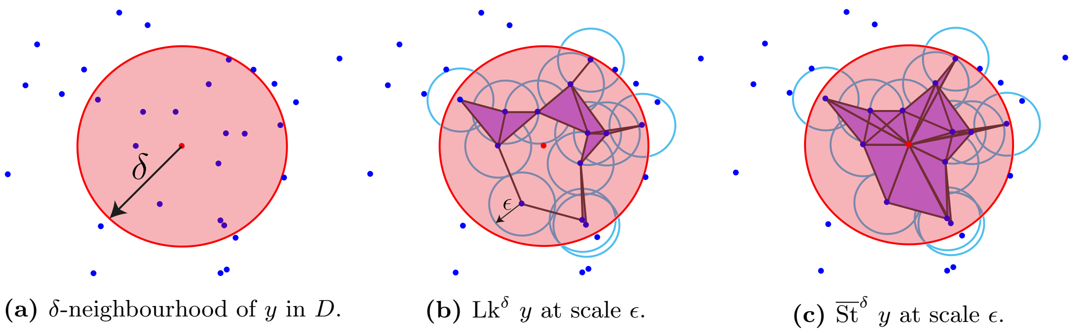
Building on the δ-link of a data point we can define the δ-star of a data point:

Definition 6 (Closed δ-star of a data point y) *Let X be a simplicial complex in a metric space constructed from a data set D, $\hat{y} \in X$ a vertex, $\delta > 0$ a distance and $\delta(\hat{y})$ a δ-neighbourhood of $\hat{y}$ in X. Then the closed δ-star $\overline{\mathrm{St}}^{\delta}\ \hat{y}$ is the union of the δ-link $\mathrm{Lk}^{\delta}\ y$ with the vertex $\hat{y}$ and all simplices $[\hat{y}, \sigma]$ where $\sigma \in X$ and σ is fully contained in $\delta(\hat{y})$.*

We note that the closed δ-star of a data point is always contractible by construction. We show an example of a data point, its δ-link and its closed δ-star in Figure 4. From now on, we refer to computing the PH of the δ-link of a point in a data set as computing the local PH of a data point. Our notion of local PH of a data point is motivated by the Mayer-Vietoris sequence and we use the property given by Equation 2 to define a new landmark selection method, in which we select points y in a data set D which are 'closest' to giving us the desired isomorphism between $PH_n(D \setminus y)$ and $PH_n(D)$ as landmarks. Note that a similar computation for local PH is performed by Wheeler et al. (2021).

3.2 Persistent Homology Landmarks

Following our local PH computations, we can now define the score $|PH_n(\mathrm{Lk}^{\delta}\ y)|$ by which we measure the failure of the isomorphism $PH_n(D \setminus y) \xrightarrow{\Psi} PH_n(D)$.



(a) δ-neighbourhood of y in D. **(b)** $\mathrm{Lk}^{\delta}\ y$ at scale ϵ. **(c)** $\overline{\mathrm{St}}^{\delta}\ y$ at scale ϵ.

Figure 4: Examples of a data point y and its δ-neighbourhood in a point cloud, the δ-link of y and the closed δ-star of y. We show the data point y in red and its δ-neighbourhood in light red highlighting the points within δ in blue. We use the data points within the δ-neighbourhood to build a Vietoris-Rips complex for a fixed filtration value $\epsilon > 0$, which represents the simplicial complex X in the definitions of the δ-link and the closed δ-star. We only show the subcomplexes of the Vietoris-Rips complex (or their extensions in the case of the closed δ-star) that are relevant to the illustrated definitions.

Definition 7 ($|PH_n(\mathrm{Lk}^{\delta}\ y)|$) *Let D be a point cloud, $y \in D$ a data point, $d : D \times D \to \mathbb{R}$ a distance function, $\Delta_y = \{\tilde{y} \in D \backslash \{y\} \mid d(\tilde{y}, y) \leq \delta\}$, $n = 0, 1, 2$ and $\mathcal{B}_n(\mathrm{Lk}^{\delta}\ y) = \{[\eta_i, \zeta_i)\}_{i=1}^{I(n)}$ the n-dimensional barcode of the Vietoris-Rips filtration performed on Δ_y excluding infinitely persisting features. Then*

$$|PH_n(\mathrm{Lk}^{\delta}\ y)| = \max_{n} \max_{i=1,\ldots,I(n)} \{\zeta_i - \eta_i\}.$$

We can re-express our observation from Equation 2 for point cloud data D and PH in the following way: if for a data point $y \in D$ and a neighbourhood radius $\delta > 0$ we can ensure that $|PH_n(\mathrm{Lk}^{\delta}\ y)| = |PH_{n-1}(\mathrm{Lk}^{\delta}\ y)| = 0$, then for $n > 0$ we obtain

$$0 \to PH_n(D \setminus y) \xrightarrow{\Psi} PH_n(D) \to 0, \tag{3}$$

where PH_n denotes the n-th homology groups associated with the different filtration steps. As we are working with data however, we are unlikely to achieve such strict conditions—indeed, we would achieve these under trivial conditions. For $|PH_n(\mathrm{Lk}^{\delta}\ y)| \approx$ small and $|PH_{n-1}(\mathrm{Lk}^{\delta}\ y)| \approx$ small, however, we obtain something close to an isomorphism between $PH_n(D \setminus y)$ and $PH_n(D)$.

The larger $|PH_n(\mathrm{Lk}^{\delta}\ y)|$ of a point is, the further away we are from an isomorphism between $PH_n(D \setminus y)$ and $PH_n(D)$. Consequently, the inclusion or exclusion of the point changes the PH of the data set more than for a point where $|PH_n(\mathrm{Lk}^{\delta}\ y)|$ is small. Under the assumption that each point has at least two neighbours within distance δ, the choice of δ determines by how much we allow the Bottleneck distance (for a definition see, for example, Otter et al., 2017) to differ between the persistence diagrams of the filtration on the point cloud containing y and a point cloud not containing y. We can therefore think of it as a resolution parameter. There are now two possible strategies for landmark selection:

- *PH landmarks I: representative landmarks.* We chose points with small $|PH_n(\mathrm{Lk}^\delta\ y)|$ values as landmarks. Their inclusion or exclusion in the full data set does not alter the PH of the full data set dramatically. Hence, the landmarks represent the data well and they do not introduce accidental topological features that can, for example, be caused by outlier points in the landmark set.
- *PH landmarks II: vital landmarks.* Landmarks are points with large $|PH_n(\mathrm{Lk}^\delta\ y)|$ values. For such points we are far away from an isomorphism between $PH_n(D\backslash\hat{y})$ and $PH_n(D)$. Hence their exclusion from the data set would change the PH dramatically.

It is not immediately clear which strategy would work better for noisy data sets. In a denoised point cloud, we would expect vital landmarks to capture the overall PH of the data well while representative landmarks could be removed from the point cloud without dramatical changes to the PH. In a noisy data set, however, noise can also introduce unwanted topological features. The removal of such noise, while strongly altering the PH of the denoised data set in comparison to the noisy version, would in such a case be desired. We therefore consider both approaches and compute $|PH_n(\mathrm{Lk}^\delta\ y)|$ for dimensions $n = 0, 1, 2$. Motivated by results from our computations, we define two versions for $|PH_n(\mathrm{Lk}^\delta\ y)|$ which we use in practice. To avoid confusion with Definition 7, we refer to these values obtained in practical computations as the *PH outlierness* of a point $y \in D$:

$$out_{\mathrm{PH}}^{0,1,2}(y) = \max\{|\mathcal{B}_0(\mathrm{Lk}^\delta\ y)|, |\mathcal{B}_1(\mathrm{Lk}^\delta\ y)|, |\mathcal{B}_2(\mathrm{Lk}^\delta\ y)|\},$$

where $|\mathcal{B}_n(\mathrm{Lk}^\delta\ y)|$ is the length of the longest finite interval in the PH barcode for dimension n. We also refer to $out_{\mathrm{PH}}^{0,1,2}(y)$ as outlierness over all (computed) dimensions. We observe that $out_{\mathrm{PH}}^{0,1,2}(y)$ is usually determined by dimension 0, where we find the longest non-infinitely persisting features in our data sets (see Subsection 5.1). To avoid this behaviour, we additionally use a version for PH outlierness which is restricted to dimension 1 in the PH calculation:

$$out_{\mathrm{PH}}^{1}(y) = |\mathcal{B}_1(\mathrm{Lk}^\delta\ y)|. \tag{4}$$

We also refer to $out_{\mathrm{PH}}^{1}(y)$ as dimension 1 outlierness. In situations where either of the two definitions can be used, we denote PH outlierness as $out_{\mathrm{PH}}(y)$.

To avoid choosing points as landmarks that are very far away from other data points, we determine points $S = \{s_1, \dots, s_o\}$ with fewer than two neighbours within their δ-neighbourhood to be *super outliers*. We assign super outliers $s \in S$ to have $out_{\mathrm{PH}}^{1}(s) = out_{\mathrm{PH}}^{0,1,2}(s) = -\infty$ and include them into the landmark set only once all other points $y \in D\backslash S$ have been chosen as landmarks. Note that the resolution parameter δ strongly influences the number of super outliers. As long as we have enough points in the data set that are not considered to be super outliers, we choose our landmarks to be the points $L = \{l_1, l_2, \dots, l_m\} \subset D \setminus S$ such that $out_{\mathrm{PH}}(l_i) \leq out_{\mathrm{PH}}(y)$ for all $y \in D \setminus \{L \cup S\}$ and $i = 1, \dots, m$ for PH landmarks I (representative landmarks) and $out_{\mathrm{PH}}(l_i) \geq out_{\mathrm{PH}}(y)$ for all $y \in D \setminus \{L \cup S\}$ and $i = 1, \dots, m$ for PH landmarks II (vital landmarks). We show the pseudocode for PH landmarks I in Algorithm 1, an algorithm for PH landmarks II can be formulated accordingly.

When using $out_{\mathrm{PH}}^{1}(y)$ in Algorithm 1, it is important to note, that one can obtain data points y with $out_{\mathrm{PH}}^{1}(y) = 0$. To avoid that the order of inclusion of such points in the

Algorithm 1 The PH landmark algorithm

```
Input: Data points $D = \{y_1, \dots, y_N\}$,
  a distance function $d : D \times D \to \mathbb{R}$
  number of landmarks $m$,
  local neighbourhood radius $\delta > 0$.
Output: A set of $m$ PH landmarks $L = \{l_1, \dots, l_m\}$, a set of $o$ super outliers
  $S = \{s_1, \dots, s_o\}$.
  for all $y \in D$ do
    Find $\Delta_y = \{\tilde{y} \in D \setminus \{y\} \mid d(\tilde{y}, y) \leq \delta\}$
    if $|\Delta_y| > 2$ then
      Compute Vietoris-Rips filtration for $\Delta_y$ for $n = 0, 1, 2$.
      Compute $out_{\mathrm{PH}}(y)$
    else
      $S \leftarrow S \cup \{y\}$
    end if
  end for
  Re-order the points in $D \setminus S$ such that $out_{\mathrm{PH}}(y_1) \leq out_{\mathrm{PH}}(y_2) \leq \cdots \leq out_{\mathrm{PH}}(y_{N-o})$
  $L \leftarrow \{y_1, \dots, y_{\min\{m, N-o\}}\}$
  if $N - o < m$ then
    $L \leftarrow L \cup \{s_1, \dots, s_{m-N+o}\}$
  end if
```

landmark set is determined by a possibly systematic and non-random ordering of the points in the original data set, which could, for example, favour noise points to be added before signal points or vice versa, we ensure that all points with $out^1_{\mathrm{PH}}(y) = 0$ are randomly permuted in the ordering of the PH outlierness scores.

3.3 $k--$ Landmarks

The k-means$--$ algorithm was developed by Chawla and Gionis (2013) to overcome the extreme sensitivity of the k-means algorithm to outliers. The authors formulate their approach as a generalisation of the k-means algorithm: for an input data set $D = \{y_1, \dots, y_N\}$ the algorithm provides a set of k cluster centres $\hat{L} = \{\hat{l}_1, \dots, \hat{l}_k\}$ and a set of j outliers $O = \{o_1, \dots, o_j\}$, $O \subset D$. For a given distance function $d : D \times D \to \mathbb{R}$ and $y \in D$ the authors use the following term in their algorithm:

$$c(y, \hat{L}) \coloneqq \arg\min_{\hat{l} \in \hat{L}} d(y, \hat{l}).$$

We show the pseudocode in Algorithm 4 in Section A.2. For our application of the algorithm to landmark selection, we further define

$$\tilde{c}(D, \hat{l}) \coloneqq \arg\min_{y \in D} d(y, \hat{l}).$$

We show our modified version of the k-means$--$ algorithm for landmark selection in Algorithm 2.

Algorithm 2 The k-means$--$ algorithm (Chawla and Gionis, 2013) modified for landmark selection

```
Input: Data points $D = \{y_1, \dots, y_N\}$, a distance function $d : D \times D \to \mathbb{R}$, number of
  clusters $k$ and number of outliers $j$.
Output: A set of $k$ cluster centers $L = \{l_1, \dots, l_k\}$,  $L \subset D$,
  a set of $j$ outliers $O = \{o_1, \dots, o_j\}$, $O \subset D$.
  $\hat{L}_0 \leftarrow \{k$ random points of $D\}$
  $e_0 = -1$
  $i \leftarrow 1$
  while  (continuation_criterion $> 10^{-4}$ and $i < 100$ ) do
    for all $y \in D$ do
      compute $d(y, \hat{l}_{i-1})$
    end for
    Re-order the points in $D$ such that $d(y_1, \hat{L}_{i-1}) \geq d(y_2, \hat{L}_{i-1}) \geq \cdots \geq d(y_N, \hat{L}_{i-1})$
    $O_i \leftarrow \{y_1, \dots, y_k\}$
    $D_i \leftarrow D \setminus O_i = \{y_{k+1}, \dots, y_N\}$Figure
    for r = 1 to k do
      $P_r \leftarrow \{y \in D_i \mid c(y, \hat{L}_{i-1}) = \hat{l}_{i-1,r}\}$
      $\hat{l}_{i,r} \leftarrow \text{mean}(P_r)$
    end for
    $\hat{L}_i \leftarrow \{\hat{l}_{i,1}, \dots, \hat{l}_{i,k}\}$
    for $y \in D_i$ do
      compute $d(y, \hat{L}_i)$
    end for
    $e_i \leftarrow \sum_{y \in D_i} d(y, \hat{L}_i)^2$
    continuation_criterion  $\leftarrow |e_i - e_{i-1}|$
    $i \leftarrow i + 1$
  end while
  $L \leftarrow \emptyset$
  $D_L \leftarrow D_{i-1}$
  while $|L| < k$ do
    for $\hat{l} \in \hat{L}$ do
      $m_{\hat{l}} \leftarrow \min_{y \in D_L} d(y, \hat{l})$
    end for
    Re-order $\{\hat{l}_1, \dots, \hat{l}_{\hat{k}}\}$ such that $m_{l_1} \leq m_{l_2} \leq \cdots \leq m_{l_{\hat{k}}}$
    $s \leftarrow 1$
    repeat
      $L \leftarrow L \cup \{y_s\}$, where $y_s = \tilde{c}(D_L, \hat{l}_s)$
      $s \leftarrow s + 1$
    until $y_s = y_t$ for some $t < s$
    $L \leftarrow L \setminus \{y_s\}$
    $D_L \leftarrow D_L \setminus L$
    $\hat{L} \leftarrow \hat{L} \setminus \{\hat{l}_1, \dots, \hat{l}_{s-1}\}$
  end while
```

3.4 Implementation

We implement PH landmark selection method in Matlab using Ripser (Bauer, 2021) for the computation of the local Vietoris-Rips complexes. We also implement the $k--$ landmarks algorithm in Matlab. For the calculation of maxmin landmarks, random landmarks, and dense core subsets we use the inbuilt functions in the javaPlex package (Tausz et al., 2014). We make an improved python version of our code for PH landmarks available on `https://github.com/stolzbernadette/Outlier-robust-subsampling-techniques-for-persistent-homology`.

4. Data Sets

We introduce the data sets to which we apply the different landmark selection methods. The data sets are chosen to be simple to allow us to determine effects of the landmark selection. The data sets consist of signal points that are sampled from a topologically interesting structure—a sphere, a Klein bottle, or a torus—and noise points that we design to be topologically different from the signal.

4.1 3-Dimensional Data Sets

We consider four 3-dimensional data sets.

4.1.1 Sphere-Cube Data Set.

For a given number of points N and probability p we sample signal points uniformly at random from the surface of the unit sphere with probability p and noise points from the (filled) cube $[-1,1]^3 \subset \mathbb{R}^3$ with probability $1-p$.

4.1.2 Sphere-Plane Data Set

For a given number of points N we sample signal points uniformly at random from the surface of the unit sphere with probability p and noise points from the xy-plane $[-3,3]^2 \subset \mathbb{R}^2$ with probability $1-p$.

4.1.3 Sphere-Line Data Set

For a given number of points N we sample signal points uniformly at random from the surface of the unit sphere with probability p and noise points from $(\alpha,0,0)$, where $\alpha \in [-50,50] \subset \mathbb{R}$, with probability $1-p$.

4.1.4 Sphere-Laplace Line Data Set

For a given number of points N we sample signal points uniformly at random from the surface of the unit sphere with probability p and we sample noise points from $(\alpha,0,0)$ with probability $1-p$, where α is sampled from $[-50,50] \subset \mathbb{R}$ and Laplace distributed with $\mu = 4$ and $\sigma = 0.5$. We use the Laplacian random number generator code (Chen, 2019) to generate α.

4.2 4-Dimensional Data Sets

We consider two 4-dimensional data sets.

4.2.1 Torus Data Set

We use the following parametrisation of the torus $\mathcal{T}$:

$$(x, y, z, \omega) = (\cos(\gamma), \sin(\gamma), \cos(\varphi), \sin(\varphi)),$$

where $\gamma, \varphi \in (0, 2\pi)$. We add noise $\mathcal{T}_{\text{noise}}$ to the torus using the equation

$$(x_{\text{noise}}, y_{\text{noise}}, z_{\text{noise}}, \omega_{\text{noise}}) = (r * \cos(\gamma), r * \sin(\gamma), \hat{r} * \cos(\varphi), \hat{r} * \sin(\varphi)),$$

where $r, \hat{r} \in (0, 2)$.

For a given number of points N we sample signal points uniformly at random from $\mathcal{T}$ with probability p and noise points from $\mathcal{T}_{\text{noise}}$ with probability $1 - p$.

4.2.2 Klein Bottle Data Set

We use the following parametrisation of the Klein bottle $\mathcal{K}$:

$$\begin{aligned} x &= \cos(\gamma) * (r * \cos(\varphi) + C), \\ y &= \sin(\gamma) * (r * \cos(\varphi) + C), \\ z &= \cos(\gamma/2) * r * \sin(\varphi), \\ \omega &= \sin(\gamma/2) * \sin(\varphi), \end{aligned}$$

where $\gamma, \varphi \in (0, 2\pi)$, $r = 3$ and $C = 2$. We define noise $\mathcal{K}_{\text{noise}}$ for the Klein bottle using the equations

$$\begin{aligned} x_{\text{noise}} &= \cos(\gamma) * (r_{\text{noise}} * \cos(\varphi) + C_{\text{noise}}), \\ y_{\text{noise}} &= \sin(\gamma) * (r_{\text{noise}} * \cos(\varphi) + C_{\text{noise}}), \\ z_{\text{noise}} &= \cos(\gamma/2) * r_{\text{noise}} * \sin(\varphi), \\ \omega_{\text{noise}} &= \sin(\gamma/2) * \sin(\varphi), \end{aligned}$$

where r_{noise} is sampled uniformly from the interval $[2, 4] \subset \mathbb{R}$ and C_{noise} is sampled uniformly from the interval $[1, 3] \subset \mathbb{R}$. We use and adapt the code from Otter et al. (2017). For a given number of points N we sample signal points uniformly at random from $\mathcal{K}$ with probability p and noise points from $\mathcal{K}_{\text{noise}}$ with probability $1 - p$.

5. Results

We present our results for the proposed landmark selection methods, i.e. PH landmarks and $k--$ landmarks. We first study PH landmark selection in detail on the sphere-cube data set with $p = 0.6$, then proceed to showing our results in comparison to the current standard methods on our various data sets, i.e. landmark selection via the maxmin algorithm and random landmark selection. We then compare our methods to dense core subsets and finally investigate the influence of the δ parameter on the performance of PH landmarks. For all landmark selection techniques we use the Euclidean distance as distance function. All of our data sets consist of 3000 points.

5.1 PH Landmarks Case Study on the Sphere-Cube Data Set with $p = 0.6$

We apply PH landmark selection with $\delta = 0.2$ to the sphere-cube data set where a data point has a probability of 0.6 to be located on the surface of the unit sphere and 0.4 to be located in the unit cube. As for all of our data sets, we find that the PH outlierness values $out_{\mathrm{PH}}^{0,1,2}(y)$, as defined in Equation 3.2, are determined by the maximal non-infinite bar in the dimension 0 barcode. We show example barcodes for dimension 0 for a noise point, a sphere point, and a super outlier in Figure 5. In general, we expect a noise point to be located in a sparser region of this data set and thus to either be classified as a super outlier, or to exhibit a barcode with a small number of long bars and very few, or no, short bars. For a sphere point, we expect the dimension 0 barcode to have many short bars and occasionally some longer bars caused by noise points that lie within the δ-neighbourhood. Note that for the examples in Figure 5, we choose the noise point with the highest outlierness score and the sphere point with the lowest outlierness score in the data set to illustrate model cases for the method. We find that for these example points, the dimension 0 barcodes behave as expected. To explore whether the outlierness scores reflect the properties of the different types of points as expected, we consider histograms of the outlierness scores $out_{\mathrm{PH}}^{0,1,2}(y)$ of all data points in Figure 6. We find that we have 48 super outliers in the data set. For the noise and sphere points, we can see that the outlier scores are distributed differently and that by including points with low outlier scores as landmarks first, we should preferentially obtain sphere points rather than noise points. Landmarks chosen in this way correspond to representative landmarks (PH landmarks I) described in Subsection 3.2.

As our landmark selection approach is motivated by Equation 3 which holds for PH in dimensions $n > 0$, it is not immediately clear that representative landmarks are preferable to vital landmarks both for $n = 0$ and dimensions $n > 0$. We examine a variant of the method where we restrict ourselves to local PH in dimension 1. In this case dimension 1 PH outlierness values correspond to the persistence of the most persistent feature in the local dimension 1 barcode (see Equation 4). We show a histogram of the distribution of the dimension 1 PH outlierness scores in Figure 7. We observe that the clearest difference between the sphere points and the noise points is that a large proportion of noise points has $out_{\mathrm{PH}}^{1}(y) = 0$, while a clear majority of the sphere points has $out_{\mathrm{PH}}^{1}(y) > 0$. This can again be explained by the fact that sphere points have more neighbours within their δ-neighbourhood and therefore are more likely to form features in dimension 1. From these observations it seems that here it is more beneficial to use vital landmarks (PH landmarks II), described in Subsection 3.2, for landmark selection based on dimension 1. For PH landmarks based on dimension 1 we thus choose points with large dimension 1 PH outlierness scores as landmarks and discard points with low dimension 1 PH outlierness scores from the data set as outliers whose removal does not alter the PH of the data set much.

5.2 Comparison of Persistent Homology Landmarks and $k--$ Landmarks to Standard Landmark Selection Methods

We compare our proposed methods for landmark selection, the $k--$ landmarks and the PH landmarks, to the current standard methods for landmark selection, i.e. random landmarks and maxmin landmarks. Using the different techniques we choose m landmarks from N data

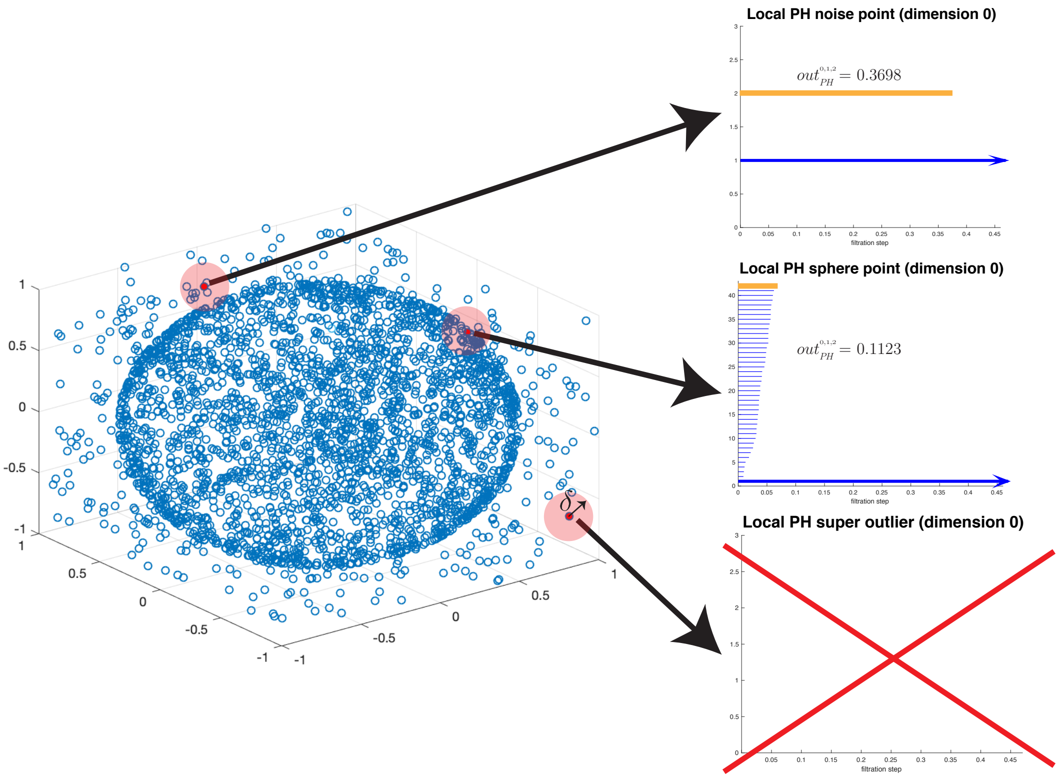


Figure 5: Schematic illustration of three different types of points y found in the sphere-cube data set, $p = 0.6$, and their local dimension 0 barcodes for $\delta = 0.2$: noise point, sphere point, and super outlier. The δ-neighbourhoods are shown in light red balls around the corresponding data points. We calculate the PH outlierness $out_{\mathrm{PH}}^{0,1,2}(y)$ for every point based on its local Vietoris-Rips barcode and ignore super outliers (indicated by the red cross drawn over the super outlier barcode). We highlight the bars in the barcodes that are used to determine $out_{\mathrm{PH}}^{0,1,2}(y)$ in yellow.

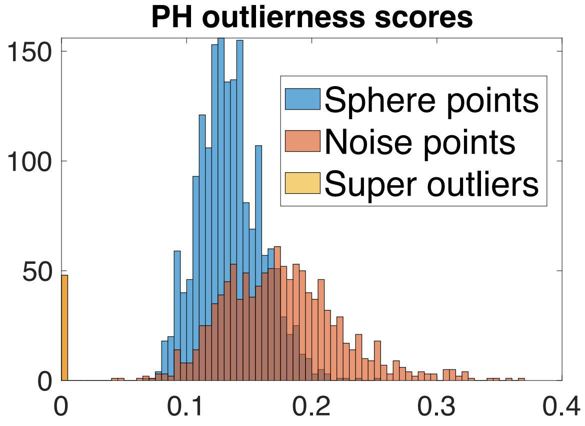


Figure 6: Histograms of the PH outlierness $out_{\mathrm{PH}}^{0,1,2}(y)$ values obtained on the sphere-cube data set, $p = 0.6$, from local PH with $\delta = 0.2$. The horizontal axis represents the outlierness scores, the vertical axis shows the number of points. Note that we assign super outliers to have PH outlierness $out_{\mathrm{PH}}^{0,1,2}(y) = -\infty$, but for illustrative purposes we present them here at $out_{\mathrm{PH}}^{0,1,2}(y) = 0$.

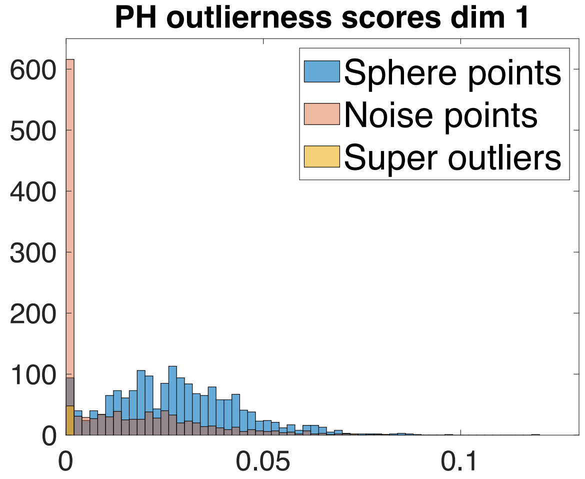


Figure 7: Histograms of the PH outlierness $out_{\mathrm{PH}}^{1}(y)$ values obtained on the sphere-cube data set, $p = 0.6$, from local PH with $\delta = 0.2$, considering only features in dimension 1. The horizontal axis represents the outlierness scores, the vertical axis shows the number of points. Note that we assign super outliers to have PH outlierness $out_{\mathrm{PH}}^{1}(y) = -\infty$, but for illustrative purposes we represent them here at $out_{\mathrm{PH}}^{1}(y) = 0$.

points. This corresponds to a landmark sampling density of $\frac{m}{N}$. For the $k--$ landmarks, we define the number of clusters to be $k = pm$ and the number of outliers to be $j = (1-p)m$, where p is the probability with which a point in the respective data set was sampled from the signal data, i.e. the sphere, torus, or Klein bottle. We consider both the case where we choose the $k--$ cluster centres and the outliers found by the algorithm as our landmarks as well as the case where we only consider the $k--$ cluster centres to be landmarks. For the PH landmarks we choose $\delta = 0.2$ for the 3-dimensional data sets, $\delta = 0.5$ for the torus data and $\delta = 0.6$ for the Klein bottle. In all our data sets, we find that when looking for the maximal persistence of a feature across all dimensions, the value of $out_{\mathrm{PH}}^{0,1,2}(y)$ is exclusively determined by dimension 0. We therefore also include a variation of PH landmarks that only considers the local dimension 1 barcode for the calculation of $out_{\mathrm{PH}}^{1}(y)$. For the PH landmark version where we use all dimensions to determine the outlierness scores, we choose data points with small $out_{\mathrm{PH}}^{0,1,2}(y)$ as our landmarks (representative landmarks). For the version where we restrict ourselves to dimension 1, we choose points with large $out_{\mathrm{PH}}^{1}(y)$ scores as landmarks (vital landmarks).

For all our data sets, our aim for the landmarks is to contain a high fraction of signal points, even when sampling only a small fraction of the data as landmarks. In Figures 8–13 we show plots of the fraction of signal points in the various landmark sets at different sampling densities. Since in the PH landmark selection we allow super outliers as landmarks once all other points are taken, we expect the fraction of signal landmarks for sampling density 1 to represent the probability of signal points in the data set, except in the variant of the $k--$ landmarks where we include only the cluster centres as landmarks (referred to as 'kMinusMinusOutlierFree' in the plots). Note that for this variant of $k--$ landmarks for data sets with $p \leq 0.5$, it is possible to obtain a signal fraction of 0 even for sampling density 1 if the algorithm selects all $k = pm$ cluster centres to be located among noise points. For the maxmin, random and $k--$ landmarks we show the average fraction of signal points and its standard deviation across 20 realisations of the selection algorithms.

As expected, we observe that the maxmin algorithm tends to select noise points as landmarks for all data sets with only one exception (see Figures 8–13): for the sphere-Laplace data, the maxmin algorithm performs well (see Figure 11), as the noise is located in a cluster far away from the signal and hence maximising the distance between landmarks results in many points being selected from the sphere. The fraction of signal landmarks for random selection also behaves as we expect for all data sets in Figures 8–13: the selected landmarks are representative for the whole data set with an almost constant fraction of signal points over all sampling densities that corresponds to the fraction of signal points in the data set. For the $k--$ landmarks, we can see a clear improvement in the signal fraction in most data sets in Figures 8—13 when considering only cluster centres as landmarks—the inclusion of outliers gives the $k--$ landmarks similarly bad properties as the maxmin landmarks. The $k--$ landmarks that do not include outliers tend to perform well for high sampling densities, where the number outlier points corresponds roughly to the number of noise points in the data set. For low sampling densities however, the method only outperforms random selection for most of the sphere-cube and Klein bottle data sets (see Figures 8 and 13). For the sphere-plane, sphere-line and the sphere Laplace-line data sets the reason for this lies in the nature of the noise, which leads to the selection of cluster centres in the noise data. We also notice large standard deviations from the average fraction of signal points in the $k--$ landmark over 20 realisations.

With exception of the sphere-line and sphere-Laplace data sets (see Figures 10 and 11), the PH landmark selection techniques I and II both outperform the standard methods as well as the $k--$ landmarks clearly for most cases, especially for low sampling densities. Interestingly, the $k--$ landmarks perform very well on the Klein bottle data set (see Figure 13), beating both PH landmarks for very small sampling densities. For the sphere-line and sphere-Laplace data sets (see Figures 10 and 11), PH landmarks II restricted to dimension 1 outperform all other methods for low sampling densities while PH landmarks I across all dimensions perform worse than all other methods in most cases. The noise in these data sets is located on lines in dense regions of the data set where the local PH does not find any topological features in dimension 1, but many features with low persistence in dimension 0. In general, the two versions of PH landmarks start coinciding as soon as super outliers are added to the data set which we can observe in the plots as a rapid drop in the fraction of signal points. We add the super outliers to the landmarks in random order (once all other points are already selected as landmarks) and hence both PH landmark methods differ only slightly in the development of their signal fractions after the addition of super outliers. We note there are cases in which the PH landmarks thrive because the points that are not super outliers are predominantly signal points. This is not the case for the sphere-line and the sphere-Laplace data sets with $p = 0.6$ (see Figures 10 and 11) for which the dimension 1 PH landmarks II perform very well. Both of these data sets have less than 4 super outliers. Interestingly, there seems to be a trend for the dimension 1 PH landmarks II to outperform PH landmarks I on data sets with high signal content, i.e. for $p > 0.6$ across all data sets. For lower signal content the PH landmarks I perform strongly. For the Klein bottle data set, dimension 1 PH landmarks II outperform PH landmarks I in almost all cases.

Overall, the results underline that both PH landmarks I and II represent the PH of the data set well and are robust to outliers, in particular for low sampling densities. They outperform standard methods in a large majority of cases and, moreover, via the PH outlierness score they give us a notion of how much a point can be considered an outlier for the respective variant of the method. $k--$ landmarks perform better than random selection in most cases and perform very well, in particular, for high sampling densities. Given the much higher computational cost and the large fluctuations in signal fraction between different realisations of the method, using random selection instead of $k--$ landmarks could however present a more practical approach.

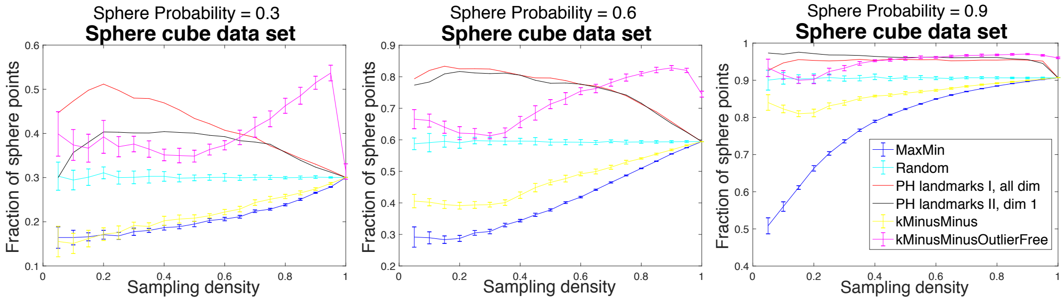


Figure 8: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-cube data set for $\delta = 0.2$. We consider landmark selection via the maxmin algorithm, random selection, PH landmarks I using $out_{\mathrm{PH}}^{0,1,2}(y)$, PH landmarks II using $out_{\mathrm{PH}}^{1}(y)$, $k--$ landmarks using both cluster centres and outliers as landmarks (kMinusMinus), and $k--$ landmarks using only cluster centres as landmarks (kMinusMinusOutlierFree).

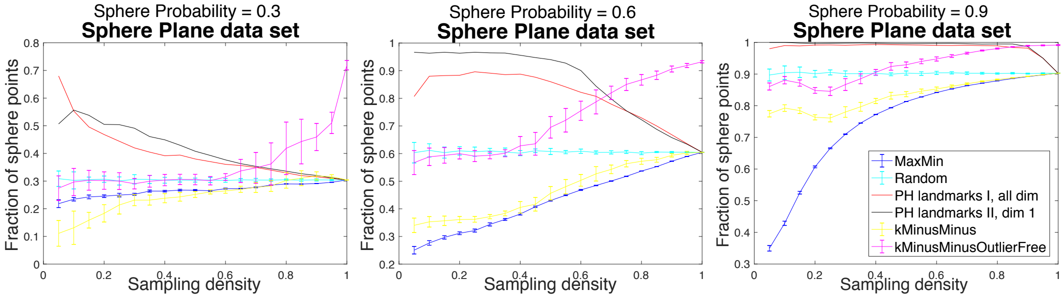


Figure 9: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-plane data set for $\delta = 0.2$.We consider landmark selection via the maxmin algorithm, random selection, PH landmarks I using $out_{\mathrm{PH}}^{0,1,2}(y)$, PH landmarks II using $out_{\mathrm{PH}}^{1}(y)$, $k--$ landmarks using both cluster centres and outliers as landmarks (kMinusMinus), and $k--$ landmarks using only cluster centres as landmarks (kMinusMinusOutlierFree).

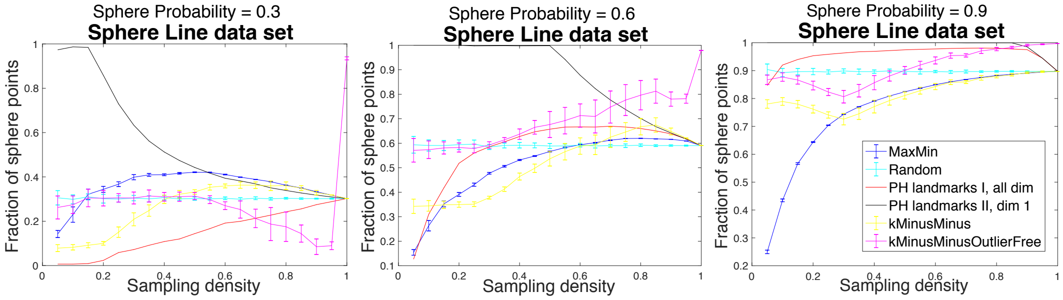


Figure 10: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-line data set for $\delta = 0.2$.We consider landmark selection via the maxmin algorithm, random selection, PH landmarks I using $out^{0,1,2}_{\mathrm{PH}}(y)$, PH landmarks II using $out^{1}_{\mathrm{PH}}(y)$, $k--$ landmarks using both cluster centres and outliers as landmarks (kMinusMinus), and $k--$ landmarks using only cluster centres as landmarks (kMinusMinusOutlierFree).

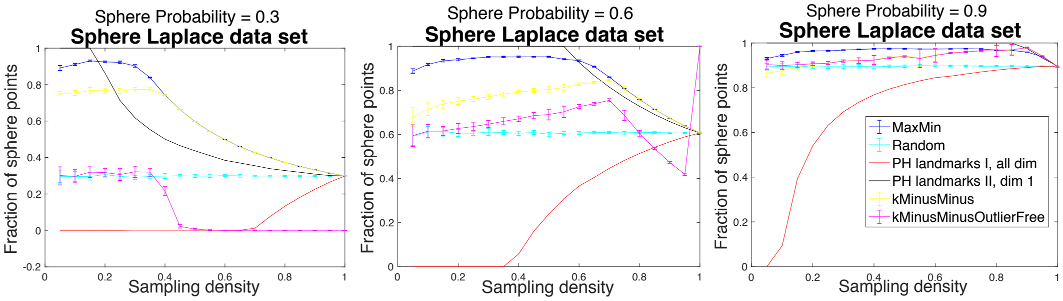


Figure 11: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the sphere-Laplace data set for $\delta = 0.2$.We consider landmark selection via the maxmin algorithm, random selection, PH landmarks I using $out^{0,1,2}_{\mathrm{PH}}(y)$, PH landmarks II using $out^{1}_{\mathrm{PH}}(y)$, $k--$ landmarks using both cluster centres and outliers as landmarks (kMinusMinus), and $k--$ landmarks using only cluster centres as landmarks (kMinusMinusOutlierFree).

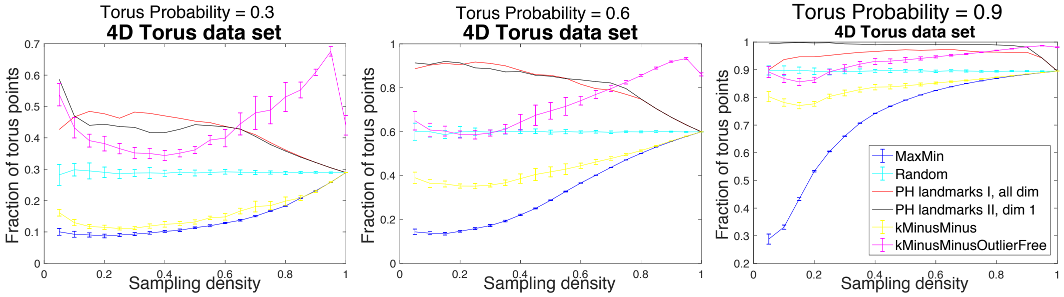


Figure 12: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the Torus data set for $\delta = 0.5$.We consider landmark selection via the maxmin algorithm, random selection, PH landmarks I using $out_{\mathrm{PH}}^{0,1,2}(y)$, PH landmarks II using $out_{\mathrm{PH}}^{1}(y)$, $k--$ landmarks using both cluster centres and outliers as landmarks (kMinusMinus), and $k--$ landmarks using only cluster centres as landmarks (kMinusMinusOutlierFree).

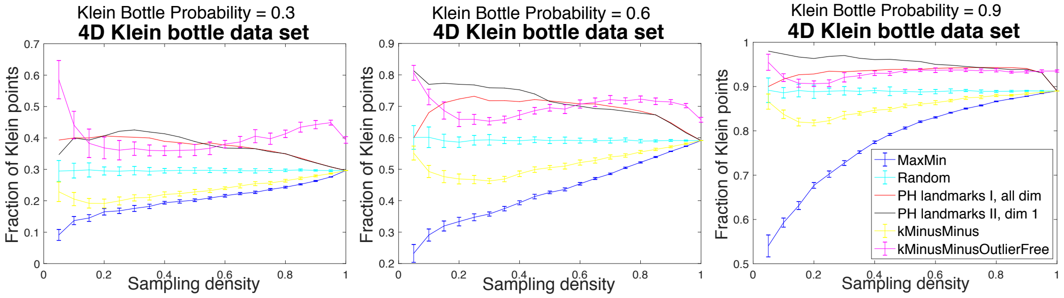


Figure 13: Comparison of the fraction of sphere points in selected landmark points for different landmark selection techniques on the Klein bottle data set for $\delta = 0.6$.We consider landmark selection via the maxmin algorithm, random selection, PH landmarks I using $out_{\mathrm{PH}}^{0,1,2}(y)$, PH landmarks II using $out_{\mathrm{PH}}^{1}(y)$, $k--$ landmarks using both cluster centres and outliers as landmarks (kMinusMinus), and $k--$ landmarks using only cluster centres as landmarks (kMinusMinusOutlierFree).

5.3 Comparisons between Persistent Homology Landmark Selection Methods and Dense Core Subsets

We now provide a more in detail study of the two PH landmark selection techniques, in particular concerning the influence of the δ parameter. We also compare the techniques to two dense core subsets, using for $K = 1$ and $K = 50$, which we consider as landmark sets. We present our results in Figures 14–19. We present only the plots for data sets with a signal probability $p = 0.6$.

For the dense core subsets, we find that for all data sets except the sphere-plane, sphere-line, and sphere-Laplace data sets (see Figures 15–17) the local density measure $K = 1$ outperforms the more global density measure $K = 50$. Indeed, in these cases, the dense core subset with $K = 1$ captures a larger fraction of signal points than most of our PH landmarks. For the sphere-cube data set (see Figure 14), we seem to outperform the dense core subset with $K = 1$ for a small range of low sampling densities for $\delta = 0.05$, which is a δ value where most of the data points are classified as super outliers. Our definition of super outliers as points with less than two neighbours in their δ-neighbourhoods, seems to imply that, for this data set, the distance to the second closest neighbour is more relevant for low sampling densities than the distance to the closest neighbour. Interestingly, for the data sets where the local dense core subset with $K = 1$ performs well (see Figures 14, 18, and 19), we also see that smaller values of δ give us better results, both for PH landmarks I across all dimensions and PH landmarks II restricted to dimension 1. For the sphere-line data set (see Figure 16), where we have a better performance for the dense core subset with $K = 50$, larger δ-neighbourhoods are more advantageous for both PH landmark versions. In this case, we perform as well as the dense core subset with $K = 50$ for $\delta = 0.3, 0.35, 0.4$ for dimension 1 PH landmarks II. In the sphere-Laplace data set (see Figure 17) PH landmarks II clearly outperform both dense core subsets. Here again, we find the trend that larger δ values are advantageous for PH landmarks II. The fact that dense core subsets perform well on most of our data sets is determined by our signal points lying in denser regions of the data than the noise points. It is only in the sphere-Laplace data set (see Figure 17), where the noise does not obey this characteristic and we observe that the local PH information in this case is richer than the distance to the K-th neighbour.

Finally, we show how the number of super outliers depends on the choice of δ in Figure 20. We observe that small δ values lead to a drastic increase in the proportion of super outliers in the data set. This underlines that, even though one can think of δ as a resolution parameter, depending on the data set, the proportion of super outliers is an important factor to consider.

Overall, PH landmarks outperform all standard techniques on most data sets as well as two dense core subsets on the sphere-Laplace for a broad range of δ-values. To obtain signal fractions that are high and stable over a long range of low sampling densities we recommend to choose δ as small as possible without having too many super outliers in the data.

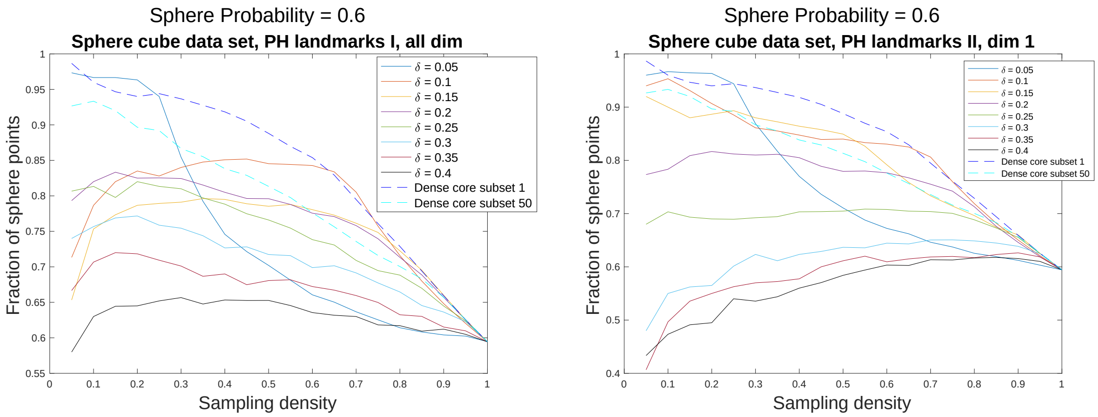


Figure 14: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-cube data set, $p = 0.6$. We show PH landmarks I using $out_{\text{PH}}^{0,1,2}(y)$ (left) and PH landmarks II using $out_{\text{PH}}^{1}(y)$ (right).

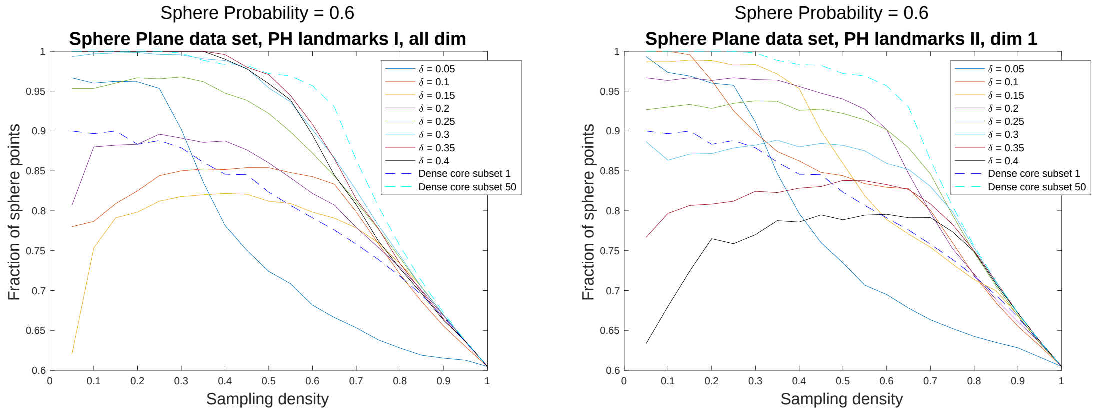


Figure 15: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-plane data set, $p = 0.6$. We show PH landmarks I using $out_{\text{PH}}^{0,1,2}(y)$ (left) and PH landmarks II using $out_{\text{PH}}^{1}(y)$ (right).

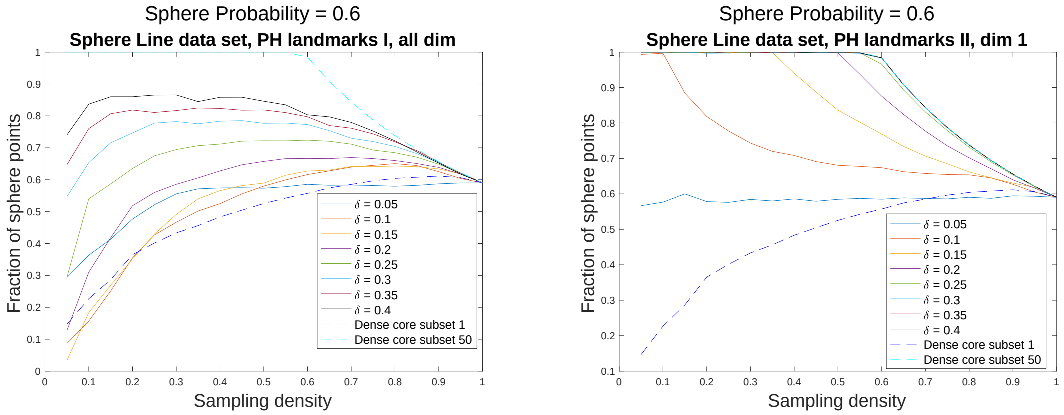


Figure 16: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-line data set, $p = 0.6$. We show PH landmarks I using $out_{\mathrm{PH}}^{0,1,2}(y)$ (left) and PH landmarks II using $out_{\mathrm{PH}}^{1}(y)$ (right).

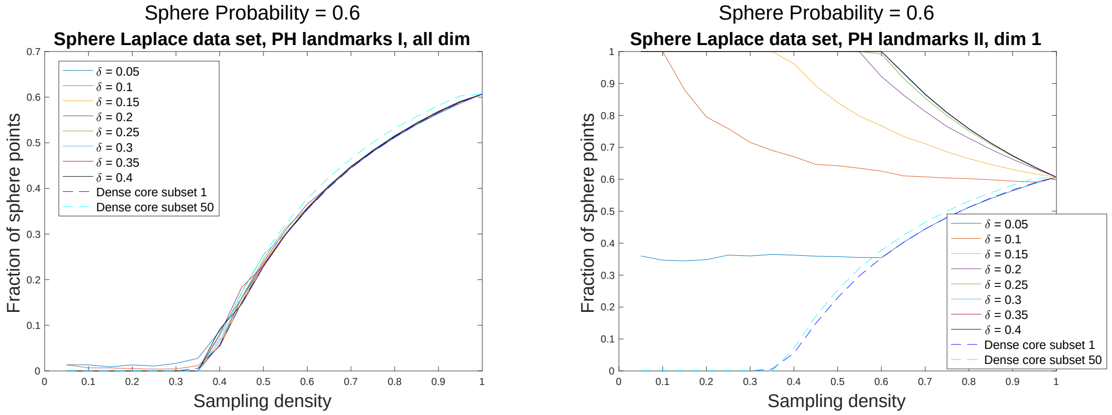


Figure 17: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the sphere-Laplace data set, $p = 0.6$. We show PH landmarks I using $out_{\mathrm{PH}}^{0,1,2}(y)$ (left) and PH landmarks II using $out_{\mathrm{PH}}^{1}(y)$ (right).

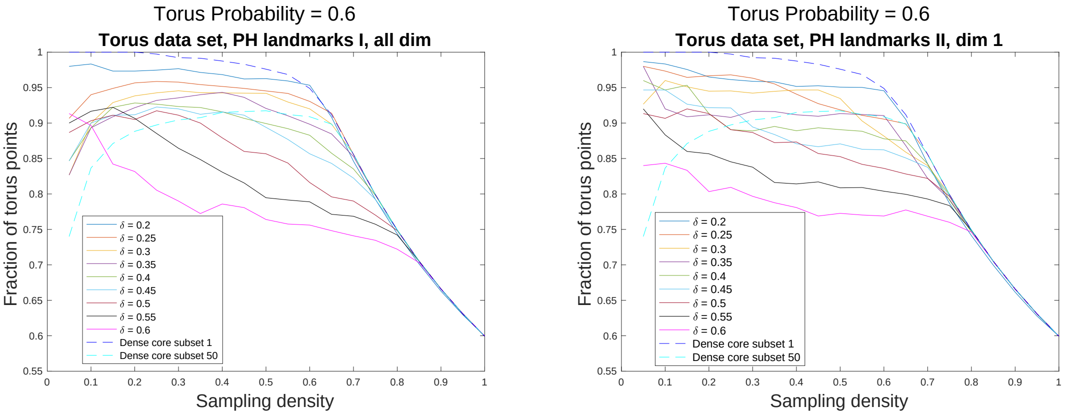


Figure 18: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the torus data set, $p = 0.6$. We show PH landmarks I using $out^{0,1,2}_{\text{PH}}(y)$ (left) and PH landmarks II using $out^{1}_{\text{PH}}(y)$ (right).

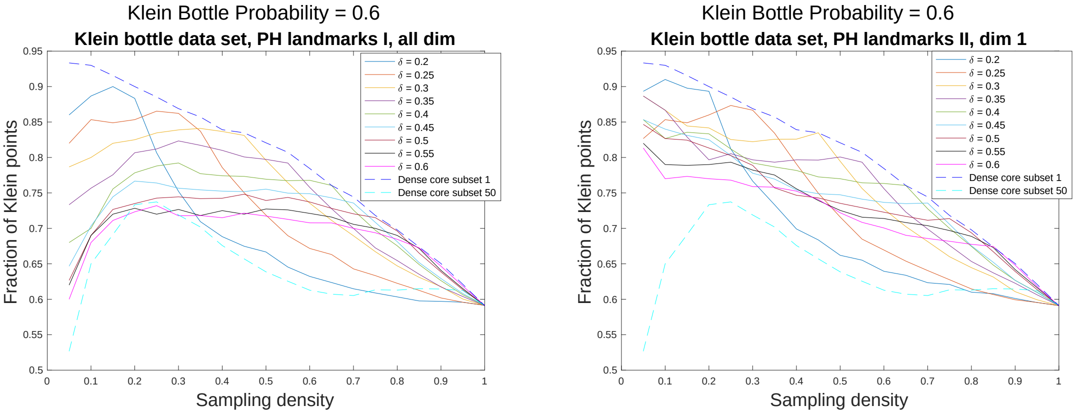


Figure 19: Comparison of the fraction of sphere points in selected PH landmark points for different values of δ and dense core subsets on the Klein bottle data set, $p = 0.6$. We show PH landmarks I using $out^{0,1,2}_{\text{PH}}(y)$ (left) and PH landmarks II using $out^{1}_{\text{PH}}(y)$ (right).

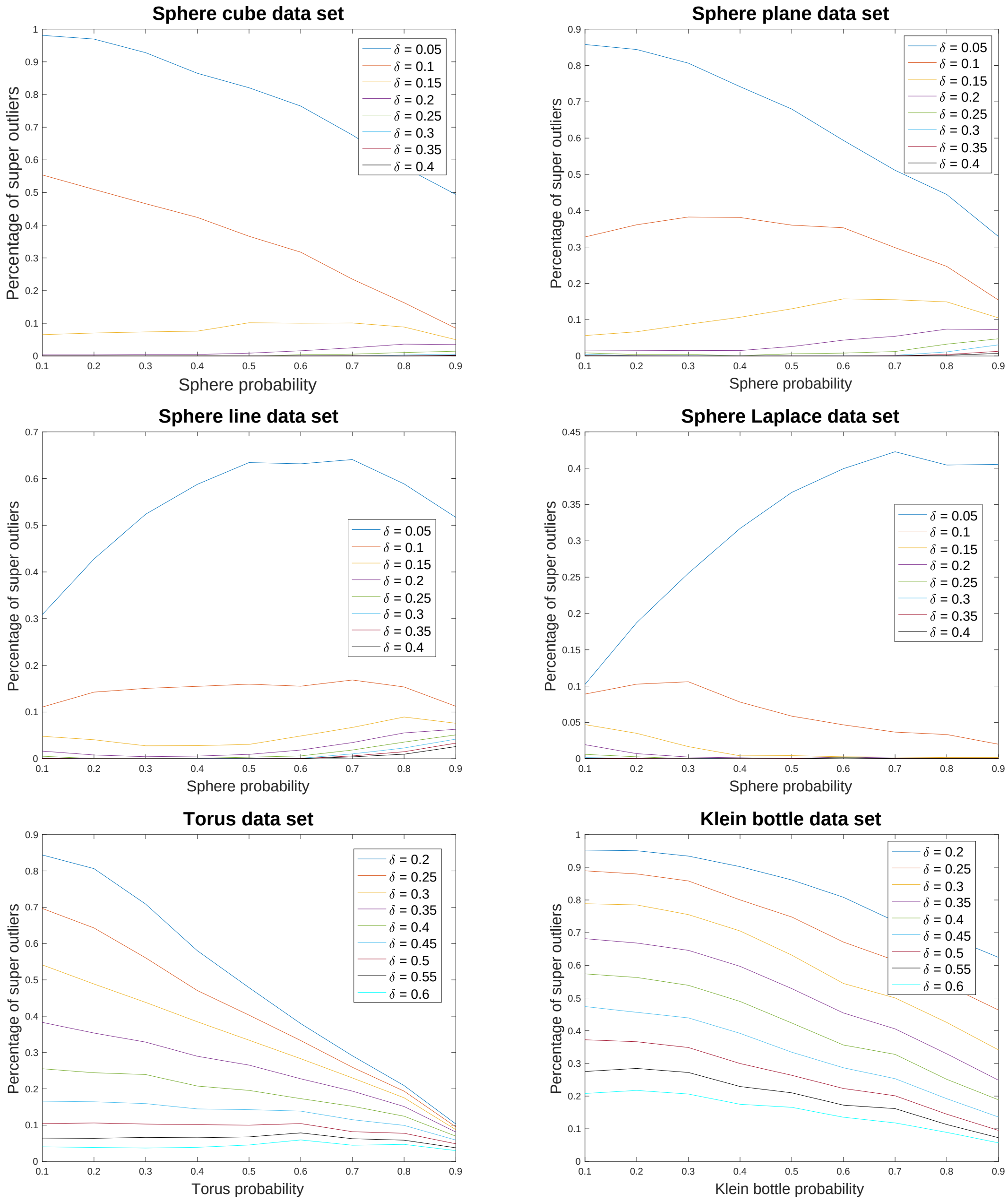


Figure 20: Influence of the choice of δ on the number of super outliers for different data sets.

6. Discussion

We proposed novel outlier-robust landmark selection techniques: $k--$ landmarks and two versions of PH landmarks. PH landmark selection is the first landmark selection method developed specifically for the use of PH.

The $k--$ landmarks outperformed existing standard landmark selection methods—random selection and maxmin selection—in many cases. They however tended to do so for large sampling densities, which are not very relevant for data subsampling. While $k--$ landmarks did meet our goals for an outlier-robust landmark selection technique, this algorithm has a high computational cost and is difficult to use on a data set with unknown properties as one has to predetermine the number of outliers for the algorithm to find.

We found that both versions of PH landmarks outperformed the existing standard landmark selection techniques on data sets containing noise, in particular for low landmark sampling densities. We observed this for a wide range of δ resolution values. In most of our data sets, PH landmarks II (vital landmarks) restricted to dimension 1 for calculating the outlierness values slightly outperformed the PH landmarks I (representative landmarks) considering all dimensions (which coincides with a restriction to dimension 0 in our cases) for data sets with a high signal content and a low noise content. We further observed that the type of noise (structured, e.g. sampled from a manifold, versus unstructured) influences the performance of the landmark selection strategy: PH landmarks I (representative landmarks) performed better for higher noise content, but typically in combination with unstructured noise. Our observations underline our interpretation of the two landmark selection strategies: for data sets with high (topological) signal content, vital landmarks (PH landmarks II) with high $out^{1}_{\mathrm{PH}}(y)$ correspond to points that have at least four local neighbours that are spread out enough to create a persistent loop in dimension 1. Removing a central point from such a neighbourhood can thus create a loop in dimension 1 and possibly contribute to holes in higher dimensions. Keeping the central point in this case is vital. The focus on dimension 1 local PH also automatically discards noise points located along lines which typically do not contribute to the overall topological structure of the data. Additionally, local PH in dimension 1 can also capture geometric anomalies (see Stolz et al., 2020, although note that in this case the computations rely on annular regions rather than δ-neighbourhoods), which can be an advantage when working with data that has structured noise, in particular if this noise has boundary regions as such points will (correctly) not be selected as vital landmarks. In contrast, for data sets with low to medium noise content, which is unstructured, representative landmarks (PH landmarks I) with low $out^{0,1,2}_{\mathrm{PH}}(y)$ are points with neighbours that are pairwise close to each other. The neighbours of representative landmarks can either be far away (within the local neighbourhood) or close to the landmark point in the centre, they can form clusters or trace out a path around the landmark point (such cases would not be captured by density measures). Unstructured noise, however, will typically not give rise to such neighbourhoods and have high $out^{0,1,2}_{\mathrm{PH}}(y)$. Thus while the noise does not have any interesting underlying topology, it can introduce topological noise, e.g. by contributing large loops between noise points and/or signal points. This type of noise - as opposed to the representative landmarks - is not representative of the topology of the signal data. We summarise our observed trends for good performance of PH landmark selection techniques depending on signal density and

Type of noise	Signal density	Better performing landmark type
Structured	High	PH landmarks II, $out^{1}_{\mathrm{PH}}(y)$
Unstructured	High	PH landmarks II, $out^{1}_{\mathrm{PH}}(y)$
Unstructured	Medium, low	PH landmarks I, $out^{0,1,2}_{\mathrm{PH}}(y)$

Table 1: Overview of landmark selection strategy performance. We observe trends for which of the two PH landmark selection strategies performs better depending on the type of noise in the data (structured, e.g. sampled from a plane or line, versus unstructured, e.g. sampled from the unit cube) and the signal density (high, i.e. approximately 90% of the data points are signal points, versus medium, i.e. approximately 60% of the data points are signal points, versus low, i.e. less than 30% of the data points are signal points). In all cases we consider low sampling density, i.e. we only include trends observed when sampling up to 50% of data points as landmarks. We restrict ourselves to cases where we observed a clear trend in landmark performance on our data sets, note that this might not generalise to other data sets.

noise type in Table 1. We hope that this summary can serve as guidance for the reader interested in applying a PH landmark selection strategy to data.

Dimension 1 PH landmarks II further outperformed dense core subsets showing that the method can capture richer information than just considering the K-th nearest neighbour as shown on the sphere-Laplace data set. On our other data sets, both versions of PH landmarks were outperformed by dense core subsets since in these data sets the signal points tend to be located in denser parts of the data than noise points. Overall, however, we consider PH landmarks to be a more practical approach than dense core subsets as they only require the choice of one parameter, δ. In contrast, when applying dense core subsets as suggested by Adams and Tausz (2015), one needs to choose the parameter K to obtain a density estimate, followed by the number of densest points to be chosen from the data set, which are then used to obtain maxmin landmarks. De Silva and Carlsson (2004) observed that when studying dense core subsets the choices of parameters can indeed result in strong topological differences in the selected point clouds. With only one parameter choice we expect both versions of PH landmarks to deliver more consistent results.

We hypothesise that PH landmarks II using $out^{1}_{\mathrm{PH}}(y)$ will lead to better results in high-dimensional data in practice, especially in cases where signal data is sparse. The parameter δ would then have to be large enough to ensure that local PH in dimension 1 produces non-trivial barcodes for a large proportion of data points. For very large data sets with potentially high noise content, calculating $out^{0,1,2}_{\mathrm{PH}}(y)$ or even a version restricted to dimension 0 for PH landmarks I could therefore be computationally advantageous.

We believe that PH landmarks contribute a valuable alternative to the current standard landmark selection techniques for PH, in particular for noisy data sets. The PH outlierness values as well as the number of super outliers with respect to the δ resolution could further be used to provide interesting insight into data sets for exploratory data analysis and could be incorporated in novel data analysis techniques which aim to take properties of shape of the data into account without performing computationally expensive PH analyses of the full data set. Identifying representative and vital landmarks in data sets could be of interest to study processes underlying the data, for example in biology. While methods from Machine Learning are excellent at classifying such data, the underlying mechanisms often remain hidden to these methods and results by themselves can be difficult to interpret. Building on our work, it would be interesting to investigate different definitions of PH outlierness in future studies. Testing the applicability of PH landmarks to real-world data sets with different types of noise or different levels of sparsity will be the subject of future work.

Acknowledgments

The author is very grateful for feedback and input from Vidit Nanda, Jared Tanner, Heather Harrington, and Lewis Marsh as well as helpful computational suggestions from Scott Balchin. BJS is a member of the Centre for Topological Data Analysis and this research was funded in whole or in part by EPSRC EP/R018472/1. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission. BJS further acknowledges a L'Oréal-Unesco For Women In Science Fellowship as well as doctoral funding from the EPSRC and MRC (EP/G037280/1) and F. Hoffmann-La Roche AG.

Appendix A. Appendix

A.1 Pseudocode for Maxmin Algorithm (Adams and Tausz, 2015)

Algorithm 3 The maxmin algorithm (Adams and Tausz, 2015)

```
Input: Data points $D = \{y_1, \dots, y_N\}$,
  a distance function $d : D \times D \to \mathbb{R}$
  number of landmarks $m$.
Output: A set of $m$ maxmin landmarks $L = \{l_1, \dots, l_m\}$.
  Select $y \in D$ at random
  $l_1 \leftarrow y$
  $L_1 \leftarrow \{l_1\}$
  $D_1 \leftarrow D \setminus \{l_1\}$
  for i = 2 to m do
    for all $y \in D_{i-1}$ do
      Calculate $d(y, L_{i-1})$
    end for
    Find $l_i$ such that $d(l_i, L_{i-1}) = \max_{y \in D} d(y, L_{i-1})$
    $L_i \leftarrow L_{i-1} \cup \{l_i\}$
    $D_i \leftarrow D_{i-1} \setminus \{l_i\}$
  end for
```

A.2 Pseudocode for k-means−− Algorithm (Chawla and Gionis, 2013)

Algorithm 4 The k-means−− algorithm (Chawla and Gionis, 2013)

```
Input: Data points $D = \{y_1, \dots, y_N\}$,
  a distance function $d : D \times D \to \mathbb{R}$,
  number of clusters $k$ and number of outliers $j$.
Output: A set of $k$ cluster centers $\hat{L} = \{\hat{l}_1, \dots, \hat{l}_k\}$,
  a set of $j$ outliers $O = \{o_1, \dots, o_j\}$, $O \subset D$.
  $\hat{L}_0 \leftarrow \{k$ random points of $D\}$
  $i \leftarrow 1$
  while (No convergence achieved) do
    for all $y \in D$ do
      compute $d(y, \hat{L}_{i-1})$
    end for
    Re-order the points in $D$ such that $d(y_1, \hat{L}_{i-1}) \geq d(y_2, \hat{L}_{i-1}) \geq \cdots \geq d(y_N, \hat{L}_{i-1})$
    $O_i \leftarrow \{y_1, \dots, y_k\}$
    $D_i \leftarrow D \setminus O_i = \{y_{k+1}, \dots, y_N\}$
    for r = 1 to k do
      $P_r \leftarrow \{y \in D_i \mid c(y, \hat{L}_{i-1}) = \hat{l}_{i-1,r}\}$
      $\hat{l}_{i,r} \leftarrow \text{mean}(P_r)$
    end for
    $\hat{L}_i \leftarrow \{\hat{l}_{i,1}, \dots, \hat{l}_{i,k}\}$
    $i \leftarrow i + 1$
  end while
```

References

Henry Adams and Gunnar Carlsson. On the nonlinear statistics of range image patches. *SIAM Journal on Imaging Sciences*, 2(1):110–117, 2009.

Henry Adams and Andrew Tausz. JavaPlex tutorial. Available at `http://javaplex.googlecode.com/svn/trunk/reports/javaplex_tutorial/javaplex_tutorial.pdf`, 2015.

Mahmuda Ahmed, Brittany Terese Fasy, and Carola Wenk. Local persistent homology based distance between maps. In *22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52. ACM, 2014.

Ulrich Bauer. Ripser: a lean C++ code for the computation of Vietoris–Rips persistence barcodes. Software available at `https://github.com/Ripser/ripser`, software retrieved in 2017.

Ulrich Bauer. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology*, 2021. doi: 10.1007/s41468-021-00071-5.

Paul Bendich, Bei Wang, and Sayan Mukherjee. Local homology transfer and stratification learning. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1355–1370. SIAM, 2012.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46: 255–308, 2009.

Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1): 1–12, 2008.

Álvaro Torras Casas. Distributing persistent homology via spectral sequences. *arXiv preprint arXiv:1907.05228*, 2019.

Sanjay Chawla and Aristides Gionis. k-means--: A unified approach to clustering and outlier detection. In *2013 SIAM International Conference on Data Mining*, pages 189–197. SIAM, 2013.

Elvis Chen. Laplacian random number generator. Available at: `https://www.mathworks.com/matlabcentral/fileexchange/13705-laplacian-random-number-generator`, March 2019.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. In Markus Gross, Hanspeter Pfister, Marc Alexa, and Szymon Rusinkiewicz, editors, *SPBG'04 Symposium on Point - Based Graphics 2004*, pages 157–166. The Eurographics Association, 2004.

Emilie Dufresne, Parker Edwards, Heather A. Harrington, and Jonathan Hauenstein. Sampling real algebraic varieties for topological data analysis. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1531–1536. IEEE, 2019.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.

Brittany Terese Fasy and Bei Wang. Exploring persistent local homology in topological data analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6430–6434. IEEE, 2016.

Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2008.

Violeta Kovacev-Nikolic. Persistent homology in analysis of point-cloud data. Master's thesis, University of Alberta, `https://era.library.ualberta.ca/files/cv43nx33b/Kovacev-Nikolic_Violeta_Fall2012.pdf`, 2012.

Svetlana Lockwood and Bala Krishnamoorthy. Topological features in cancer gene expression data. In *Pacific Symposium on Biocomputing*, pages 108–119, 2015.

Yuriy Mileyko. Another look at recovering local homology from samples of stratified sets. *Journal of Applied and Computational Topology*, 5(1):55–97, 2021.

James R. Munkres. *Elements of Algebraic Topology.* The Benjamin/Cummings Publishing Company, inc., Redwood City (California), Menlo Park (California), Reading (Massachusetts), Amsterdam, Don Mills (Ontario), Mexico City, Sydney, Bonn, Madrid, Singapore, Tokyo, Bogota, Santiago, San Juan, Wokingham (United Kingdom), 1984.

Vidit Nanda. Local cohomology and stratification. *Foundations of Computational Mathematics*, 20(2):195–222, 2020.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1-3):419–441, 2008.

Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. A roadmap for the computation of persistent homology. *European Physical Journal – Data Science*, 6(17):1–38, 2017.

Michael Robinson, Chris Capraro, Cliff Joslyn, Emilie Purvine, Brenda Praggastis, Stephen Ranshous, and Arun Sathanur. Local homology of abstract simplicial complexes. *arXiv preprint arXiv:1805.11547*, 2018.

Gurjeet Singh, Facundo Mémoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L. Ringach. Topological analysis of population activity in visual cortex. *Journal of Vision*, 8(11):1–18, 2008.

Primoz Skraba and Katharine Turner. Wasserstein stability for persistence diagrams. *arXiv preprint arXiv:2006.16824*, 2020.

Bernadette J. Stolz, Jared Tanner, Heather A. Harrington, and Vidit Nanda. Geometric anomaly detection in data. *Proceedings of the National Academy of Sciences*, 117(33): 19664 – 19669, 2020. ISSN 0027-8424. doi: 10.1073/pnas.2001741117. URL `https://www.pnas.org/content/early/2020/07/31/2001741117`.

Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. JavaPlex: A research software package for persistent (co)homology. In Han Hong and Chee Yap, editors, *Mathematical Software. ICMS 2014*, volume 8592 of *Lecture Notes in Computer Science*, pages 129–136. Springer, Berlin, Heidelberg, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

Matthew Wheeler, Jose Bouza, and Peter Bubenik. Activation landscapes as a topological summary of neural network performance. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3865–3870. IEEE, 2021.