Natural computation techniques for uncovering low-dimensional topological structures in large scale astronomical simulations

Taghribi, Albolfazl

DOI:
10.33612/diss.250007790

Publication date:
2022

Based on:

A. Taghribi, M. Canducci, M. Mastropietro, S. De Rijcke, K. Bunte, P. Tino – *"ASAP - A Sub-sampling Approach for Preserving Topological Structures Modeled with Geodesic Topographic Mapping,"*. Neurocomputing, vol. 470, pp. 376-388, 2022. Available: 10.1016/j.neucom.2021.05.108.

A. Taghribi, K. Bunte, M. Mastropietro, S. De Rijcke, P. Tino – *"ASAP : A Sub-sampling Approach for Preserving Topological Structures,"*. ESANN 2020, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Proceediings. Ciaco - i6doc.com; 2020. p. 67–72.

Chapter 3

ASAP - A Sub-sampling Approach for Preserving Topological Structures Modeled with Geodesic Topographic Mapping

Abstract

Topological data analysis tools enjoy increasing popularity in a wide range of applications, such as Computer graphics, Image analysis, Machine learning, and Astronomy for extracting information. However, due to computational complexity, processing large numbers of samples of higher dimensionality quickly becomes infeasible. This contribution is two-fold: We present an efficient novel sub-sampling strategy inspired by Coulomb's law to decrease the number of data points in d-dimensional point clouds while preserving its homology. The method is not only capable of reducing the memory and computation time needed for the construction of different types of simplicial complexes but also preserves the size of the voids in d-dimensions, which is crucial e.g. for astronomical applications. Furthermore, we propose a technique to construct a probabilistic description of the border of significant cycles and cavities inside the point cloud. We demonstrate and empirically compare the strategy in several synthetic scenarios and an astronomical particle simulation of a dwarf galaxy for the detection of superbubbles (supernova signatures).

3.1 Introduction

Topological data analysis (TDA) provides exploration tools for increasingly diverse applications in various domains, ranging from Biology and medical images (Dey, Hou and Mandal 2019), mapping disease spaces (Torres et al. 2016), and Astronomy (Xu et al. 2019). Persistent homology (PH) is a TDA technique for computing the properties of shapes of a finite metric space (also called point cloud dataset) and can capture these features in an extended range of scales. Nonetheless, as the number of points or the dimensions of a dataset increases, the computation of the PH soon becomes impractical.

Numerous methods and toolboxes provide novel approaches to tackle the problem of computational costs. Sparse Rips filtration (Sheehy 2013) builds an ϵ-net on top of the point set followed by an association of weights to each node, which results in a provably good approximation of the full data Rips filtration. In (Dey et al. 2014) two new atomic operations for efficient computation of PH are suggested, and SimBa (Dey, Shi and Wang 2019) combines these two strategies to reach a higher sparsity in the number of simplices, which increases the efficiency for computation of Rips filtration. The toolbox Ripser (Bauer 2019) decreases the computational costs by avoiding to build the complete coboundary matrix building and storing only the parts needed. This improves the memory consumption and reduces the computational time. These methods are limited to Rips and are not extendible to other types of filtration.

A general concept for scaling down the computation independent of the filtration was reported in (Chazal et al. 2015) proposing to sub-sample the data randomly repeatedly and construct an average landscape for the point cloud. Although their approach can be applied for constructing all types of filtration, it is sensitive to the distribution of the data on the structures as a consequence of random sampling. MaxMin (De Silva and Carlsson 2004) was introduced as another intuitive sub-sampling approach. By selecting a random data point as the first sample, it continuously picks the next sample point that has the longest distance to the previous samples until the desired number of samples is achieved. Although sampling using this method usually achieves more uniformly spaced distribution of points than random sub-sampling (Chazal et al. 2015), it does not provide any information about the range or distance between samples, and final results may vary a lot dependent on the starting point.

When applying a filtration to a point set, topological features appear and disappear (referred to as birth and death) by increasing the filtration parameter value. Topological features exhibiting a short lifetime are considered as topological noise in some applications, as explained in (Fasy et al. 2014). They introduced the confidence band inspired by the p-value definition from statistics. Using this definition, one can distinguish between properties which belong to the point cloud and do not emerge as artifacts due to the sub-sampling of the data. Furthermore, the persistence diagram does not provide any information about the location of these features inside the point cloud. This location information is essential in several applications, such as medical image segmentation (Dey, Hou and Mandal 2019), detecting voids in the cosmic web (Xu et al. 2019) and supernovae in galaxies. In (Dey, Hou and Mandal 2019) a technique for positioning persistent 1-cycles was introduced, which is not easily extendable for locating cavities and higher dimensional properties. Moreover, Dionysus (*Dionysus, a C++ library for computing persistent homology* n.d.) can also record the boundaries of a topological feature during the computation of PH. In (Xu et al. 2019), the authors use this toolbox to locate the voids and filaments in the Cosmic web. The recovered boundary, however, is often not fully located on the border of a hole or cavity and varies with repeated sampling over the point cloud. They furthermore construct the filtration on top of a 3D grid and then compute the distance-to-measure function (Chazal et al. 2018) for every point on the grid. As a consequence the boundary points also fluctuate by changing the grid size. We will discuss the above mentioned problems in detail in section 3.2.3.

While simplicial complexes and filtrations are useful for producing clean representations of noise-free data sets, they are not as effective when applied to intrinsically noisy structures. In these cases, a probabilistic description of the low dimensional structures is desirable, as a way to capture the underlying nature of the observed data. Existing techniques are for example non-parametric density estimators, such as Parzen windows (Parzen 1962), its extensions Manifold Parzen Windows (Vincent and Bengio 2003) and Fast-Parzen Windows (Wang et al. 2009) or semi-parametric generative models like the Infinite Gaussian Mixture Model (Rasmussen 1999). However, despite fitting observed points with high accuracy those techniques are blind toward the low-dimensional nature of the structures and often the computational costs for training and evaluation is prohibitive. As an alternative, Generative Topographic Mapping (GTM) (Bishop et al. 1998b) models a noisy manifold as a low dimensional, linear, latent space embedded in the ambient space through a non-linear mapping function. The corresponding noise is defined as a Multivariate Gaussian Mixture Model (GMM) (Bishop 2006), with centers constrained to lie on the embedded latent space. However, despite the non-linearity of the mapping function, classical GTM is insufficiently flexible to model cavities and holes, which are non homeomorphic to a linear subpsace.

Physical particle simulations are one way of investigating astronomical phenomena such as galaxies and supernovae. Radiation and winds from massive stars at the end of their life can greatly affect the dynamics of gas in the interstellar medium (ISM) and in turn, change the structure of the galaxy and its ability to create new stars. Dwarf galaxies are very sensitive to the physical processes determining their evolution due to their low mass and are therefore used as probes to characterize, study and isolate them in simulations. Similar to real dwarfs simulated irregular galaxies have a very clumpy ISM and holes due to supernovae visible in the gas density distribution (Zhang et al. 2012, Verbeke et al. 2017). The characterization of the distribution of supernova shells in the ISM (so-called superbubbles), and the energies of the expanding shells (Oey and Clarke 1997a, Stanimirovic 2006), can shed light on the feedback physical processes. Superbubbles are of great astronomical interest but typically measured by eye in available catalogues and automatic tools are highly desirable.

The main contribution of this chapter is A Sub-sampling Approach for Preserving topological structures ASAP[1] (Taghribi et al. 2020), that reduces the computational cost suitable for different types of PH filtration on a general d-dimensional point clouds, for a large number of samples. In this chapter the structures found by subsequently performed PH are statistically analyzed to determine their robustness. Additionally, we propose a strategy to provide a probabilistic description of the shell of these bubbles, which, in our astronomical application, provides additional information about the supernovae borders and the stars that shape these borders. In order to fully capture the properties of such cavities, taking advantage of their low dimensional nature, we propose a modified version of the GTM: geodesic GTM (gGTM). Through this formulation, the topological features of the modelled structures are accounted for by embedding a closed low dimensional latent space onto the ambient space of the point cloud. Through the new latent space formulation we are finally able to interpret the topological structure of manifolds, embedded in higher dimensional spaces, while still capturing their natural stochasticity.

In the following, the novel sub-sampling strategy, statistical analysis, and probabilistic description is explained in detail. We then compare to state-of-the-art methods in several controlled experiments and finally investigate a snapshot of an astronomical particle simulation by computing the number and size of superbubbles within a jelly-fish like dwarf galaxy.

3.2 Methods

This section consists of three main parts. First, we describe the sub-sampling procedure followed by the calculation of the confidence band on the PH plot and its interpretation. Every time the point cloud is sampled, we extract the boundaries of significant features in the PH plot. Finally, as the boundary points fluctuate between samples, we suggest a probabilistic description of the border of cycles and cavities in the PH plot.

[1] The source code and synthetic datasets are publicly available at https://github.com/abst0603/ASAP

3.2.1 The sub-sampling approach ASAP

Computing the PH for the analysis of the evolution of shapes across different resolutions is often prohibitive due to the combinatorial nature of existing algorithms complexity, in both time and space. Therefore, we propose a two-stage strategy based on sub-sampling and Coulomb's law (Halliday et al. 2013). As described before, we first sub-sample points from the point cloud data set N (finite metric space) to reduce the amount of computation time and memory. The subset $N_r \subset N$ aims to contain fewer points $\boldsymbol{s} \in N_r$ for which the persistence diagram $D(N_r)$ approximates the persistence diagram $D(N)$ of the full data. Therefore the set N_r has to satisfy the following two conditions (Sheehy 2013) checked in every step:

(1) covering $N_r = \{\forall \boldsymbol{p} \in N, \exists \boldsymbol{s} \in N_r | \, d(\boldsymbol{p}, \boldsymbol{s}) \leqslant r\}$ and
(2) packing $d(\boldsymbol{s}_i, \boldsymbol{s}_j) > r \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in N_r$ with $i \neq j$.

We satisfy (1) by selecting a random point $\boldsymbol{s}_i$, insert it to N_r and remove all points $\{\boldsymbol{p}_j\}$ from N belonging to an open ball centered around $\boldsymbol{s}_i$ with radius r:

$$\begin{aligned} & B(\boldsymbol{s}_i, r) = \{\boldsymbol{p} \in N : d(\boldsymbol{s}_i, \boldsymbol{p}) \leqslant r\} \\ & \Rightarrow N \leftarrow N \backslash (B(\boldsymbol{s}_i, r) \cup \{\boldsymbol{s}_i\}) \text{ and } N_r \leftarrow N_r \cup \{\boldsymbol{s}_i\}. \end{aligned} \tag{3.1}$$

The process is repeated until the point set N is depleted, implicating that all points are covered by at least one open ball of a sample point in N_r. Due to the removal of points in every step, the packing condition is also fulfilled for all remaining points with distance larger than r from $\boldsymbol{s}_i$ in N.

The sub-sampling strategy fulfils both necessary conditions, but the result is not completely uniform, and the pairwise distance of any sample point pair is between r and $2r$. However, it is more desirable to have sample points equidistant from each other forming a uniform grid. As a result, we expect when all points on its boundary connect to each other it coincides with the birth time of the void. Moreover, in astronomical applications it is crucial to measure the size of the cycles, cavities and streams as accurately as possible, for which N_r needs to contain the borders of the data. Therefore we propose an extension to the sampling inspired by the movement of identical electrical particles, such as electrons, on the surface of a conductive sphere (Halliday et al. 2013). The electrons will repel each other based on Coulomb's law and approximate a uniform distribution. To take advantage of this physical repulsion force each sample is repelled by neighbouring samples by

$$\boldsymbol{m}_i = \text{disp}(\boldsymbol{s}_i) = \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} \frac{\boldsymbol{s}_j - \boldsymbol{s}_i}{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|} \cdot \frac{\gamma}{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|^2} \quad , \tag{3.2}$$

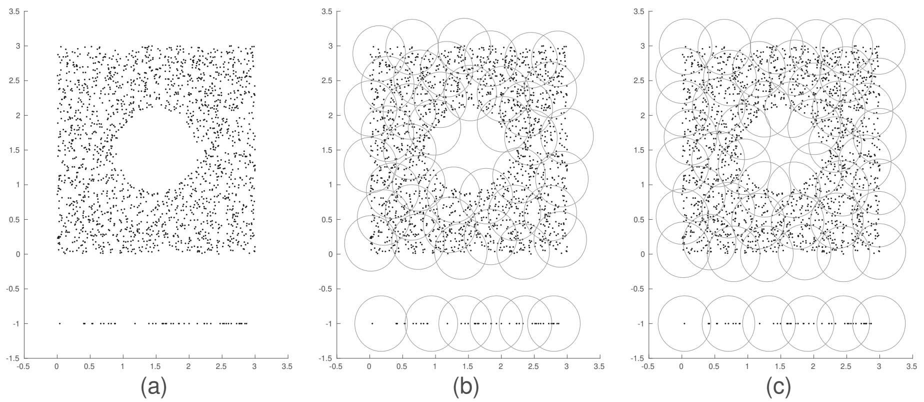

Figure 3.1: (a) points N distributed on a line and holed square, (b) ball cover after random sub-sampling and (c) after repulsive selection.

where the set $\mathcal{N}_i$ consists of sample points in $2r$ radius of $\boldsymbol{s}_i$ and γ denotes the learning rate. If neighbouring points are far from $\boldsymbol{s}_i$ the force will be low, and the learning rate controls the strength of the movement. The appropriate range for the displacement is between $(0.1r, r)$, since the effect of smaller movements is negligible and larger movements result in $\boldsymbol{s}_i$ intruding positions already covered by other samples. The learning rate is gradually reduced in every step t following

$$\gamma = r^3 \exp(-t/\tau) \ , \tag{3.3}$$

such that the samples converge to the new positions. τ is a constant which determines the decay rate of the learning rate. Instead of moving the samples itself we take the closest point in the original set $\hat{\boldsymbol{s}}_i \in N$ to the displacement position as substitute for $\boldsymbol{s}_i$

$$\hat{\boldsymbol{s}}_i = \arg\min_{\boldsymbol{p}_j}(d(\boldsymbol{p}_j, \boldsymbol{s}_i + \boldsymbol{m}_i)) \quad \forall \boldsymbol{p}_j \in N \tag{3.4}$$

if it is not contained in an open ball of any other sample point. Algorithm 3 details the complete procedure of the extended sampling strategy and Figure 3.1 shows the result on a simple two-dimensional example. Panel (a) depicts the point cloud N consisting of a line and a square with a circular hole in the centre and (b) shows the open ball cover after random sampling. The balls of N_r after the update using the repulsion force are illustrated in (c) achieving a more uniform grid that covers all boundaries as desired.

The computational complexity of Algorithm 3 depends on the number of times that *repulsion forces* are iterated, which depends on the data and the learning rate.

Algorithm 3: ASAP a sub-sampling approach preserving topological structures

```
Input  : data $N$, radius $r$, learning rate constant $\tau$
Output: $N_r$
initialise: $N_{\text{tmp}} = N$, $N_r = \emptyset$, $\gamma = 1$, and $t = 1$
while ($N_{\text{tmp}} \neq \emptyset$)
    Select a random point $\boldsymbol{s}_i$ from $N_{\text{tmp}}$
    $N_r \leftarrow N_r \cup \{\boldsymbol{s}_i\}$ and remove points from $N_{\text{tmp}}$ following Eq. (3.1)
while ($\gamma > 0.1r^3$)                                        /* repulsion forces */
    Calculate $\gamma$ based on Eq. (3.3)
    forall ($\boldsymbol{s}_i \in N_r$)
        Compute $\boldsymbol{m}_i$ Eq. (3.2) and $\hat{\boldsymbol{s}}_i$ using Eq. (3.4)
        if ($d(\hat{\boldsymbol{s}}_i, \boldsymbol{s}_j) > r \ \forall \boldsymbol{s}_j \in N_r$ AND $\boldsymbol{s}_j \neq \boldsymbol{s}_i$)
            $\boldsymbol{s}_i = \hat{\boldsymbol{s}}_i$
    $N_{\text{tmp}} = N$
    forall ($\boldsymbol{s}_i \in N_r$)                            /* fulfil covering condition */
        Remove all points belonging to $B(\boldsymbol{s}_i, r)$ from $N_{\text{tmp}}$
    while ($N_{\text{tmp}} \neq \emptyset$)
        Select a random point $\boldsymbol{s}_i$ from $N_{\text{tmp}}$
        $N_r \leftarrow N_r \cup \{\boldsymbol{s}_i\}$ and remove points from $N_{\text{tmp}}$ following Eq. (3.1)
    $t++$
```

With our choice of learning rate we typically observe about 10 iterations in our experiments. Formally, assuming the while loop on repulsion forces iterates k times in the d-dimensional data set that contains $|N|$ points over the maximum number of samples denoted by Δ_s, the worst case complexity can be written as $\mathcal{O}(kd|N|\Delta_s)$. Here we discuss the general case, however, in our implementation we employ k-d trees (Blanco and Rai 2014) for the neighborhood search in Eq. (3.1), (3.2) and (3.4), which reduces the computational complexity for pairwise distances from squared to log linear.

3.2.2 Confidence bands of significant features

The persistence diagram illustrates the birth and death time of topological features for a unique point cloud. These features in every dimension represent a specific property of the dataset, such as connected components (H_0), holes (H_1), cavities (H_2), etc. As a result, the derived persistence diagram of sampled data $D(N_r)$ does not entirely resemble the persistence plot of the point cloud $D(N)$. The Bottleneck distance is a metric of comparing two persistence diagrams (Boissonnat et al. 2018), and it is defined as follows

$$d_B(D(N), D(N_r)) = \inf_{\mu: D(N) \to D(N_r)} \sup_{\tilde{p} \in D(N)} \|\tilde{p} - \mu(\tilde{p})\|_\infty \ . \tag{3.5}$$

Here μ is a bijection that maps every feature point $\tilde{p}$ of $D(N)$ to a point on $D(N_r)$. The diagonal line where the birth and death time of features are identical is assumed to include an infinite number of points such that if the number of feature points in the persistence diagram of N and N_r is not the same, the extra points are paired with the points on the diagonal line.

Since the persistence diagram varies for distinguished sets of samples, a confidence band was introduced to separate significant topological features from noise (Fasy et al. 2014). To this aim, we follow the bootstrap procedure as described by (Fasy et al. 2014). However, we either sample the data based on Random Subsampling Method (Chazal et al. 2015) (abbreviated by RSM in the following), MaxMin (De Silva and Carlsson 2004), or our novel method ASAP. Next, for a given significance level $\alpha \in [0, 1]$, we determine c_n such that

$$\lim_{n\to\infty} \sup P(d_B(D(N), D(N_r)) > c_n) \leqslant \alpha \ . \tag{3.6}$$

As a result $C_n = [0, c_n]$ is an asymptotic $(1 - \alpha)$ confidence set for the bottleneck distance $d_B(D(N), D(N_r))$. A $(1 - \alpha)$ confidence set determines the region on a persistence diagram where we detect topological signal in the dataset with $1 - \alpha$ confidence. According to (Fasy et al. 2014), a confidence band could be included in persistence diagram with a band width of $\sqrt{2}c_n$, which specifies if a point in diagram should be considered as signal or noise.

3.2.3 Locating cavities and cycles

Following distinguishing significant topological signals and measuring their radius size, the other on-demand information is to identify and describe the point set that builds the observed feature. For instance, as explained in the introduction, a decisive step in observing a supernova in a simulated galaxy is to describe its shell or boundary. To this aim, we took advantage of the Dionysus toolbox (*Dionysus, a C++ library for computing persistent homology* n.d.), which also records the location of the topological feature generator during the computation of the persistence diagram. However, this method returns the boundaries of cycles or cavities that even may contain points outside the border of the structure. Figure 3.2(a) exemplifies one cycle boundary detected by (*Dionysus, a C++ library for computing persistent homology* n.d.), and indeed the boundary of the cycle misses some border points of the hole and invades the structure, even in this ideal situation.

One way of overcoming this problem is to locate the boundary of the same cycle in every taken sub-samples during the bootstrap procedure. Consequently, all parts of the border of a hole are recorded through sampling and locating the same hole several times. We collect the boundary points retained through such repeated sampling in a multiset Γ. The set of distinct points from Γ form the set $\bar{\Gamma}$. Multiplicity $m(\boldsymbol{b})$ of each boundary point $\boldsymbol{b} \in \bar{\Gamma}$ expresses how many times $\boldsymbol{b}$ was selected in a bootstrap procedure during the repeated sampling by ASAP.

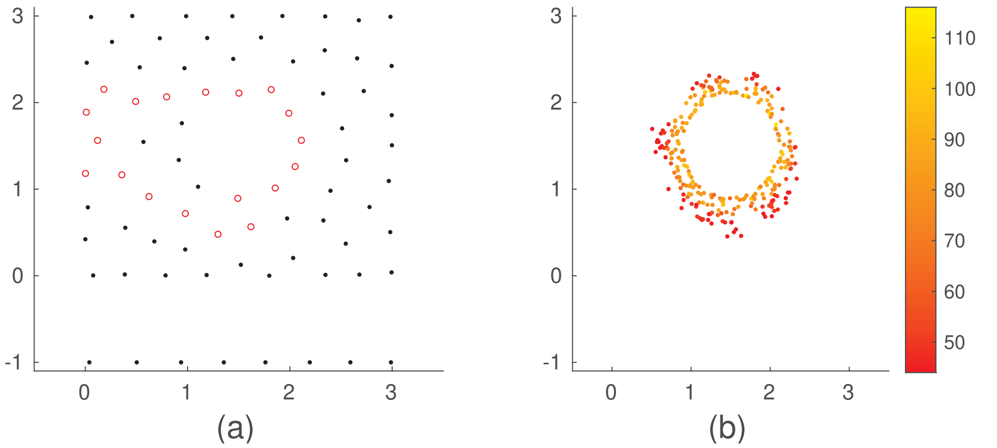

Figure 3.2: (a) Sample points extracted by ASAP (black) and the boundary of the cycle found by (*Dionysus, a C++ library for computing persistent homology* n.d.) (red). Panel (b) corresponds to the border recovered by the voting system (100 runs) with the colorbar depicting the number of votes.

To stabilize the selection of boundary points, we created the following voting scheme: First a tolerance ball $B(\boldsymbol{b}, r)$ of radius r is created around every $\boldsymbol{b} \in \bar{\Gamma}$. Next, a collection of counter variables n_j^i for $\boldsymbol{b}_i, \boldsymbol{b}_i \in \bar{\Gamma}$ is constructed, such that $n_i^j = m(\boldsymbol{b}_i)$ if $\boldsymbol{b}_i \in B(\boldsymbol{b}_j, r)$, otherwise $n_i^j = 0$. Therefore the vote for a potential boundary point $\boldsymbol{b}_i \in \bar{\Gamma}$ is computed as

$$v(\boldsymbol{b}_i) = \sum_{\boldsymbol{b_j} \in \bar{\Gamma}} n_i^j \ .$$

Finally, we sort the points by their vote value and save the ones with the upper quartile v_3 of the votes. Figure 3.2(b) shows the result of the voting operation (100 times) on the single cycle inside the data that recovers the border very satisfactory.

3.2.4 Building probabilistic models of cavities

Having identified the cavities, we would now like to capture their shape and location in the form of a probability density model aligned with the individual cavities sampled by points $\boldsymbol{b}_i$. Given the low dimensional nature of such topological features, we will model the probability distribution as the Generative Topographic Mapping (GTM) (Bishop et al. 1998b). The main idea behind GTM is to represent inherently low-dimensional structures (manifolds) embedded in a higher dimensional space by constructing a mapping via Radial Basis Functions (RBF) from a linear latent space in $\mathbb{R}^\ell$ ($\ell > 0$ being the intrinsic dimension of the manifold) to the embedding space $\mathbb{R}^D$. With classical GTM, the resulting embedded manifold is always a "stretched" and "bent" version of the linear latent space. While such a model has been shown to be very useful in capturing densities aligned along nonlinear embeddings of linear spaces (e.g. deformed "sheets of paper" embedded in $\mathbb{R}^D$ with $D \geqslant 3$) (Tino and Nabney 2002), it cannot naturally capture closed manifolds such as cycles and spheres. To make GTM applicable to our case we need to modify the latent space definition. The resulting density modelling algorithm is in the following referred to as "geodesic GTM (gGTM)".

Let us concentrate on the case of holes with their corresponding spherical latent space. Consider the sphere centered at $\mathcal{O} = (0, 0, 0) \in \mathbb{R}^3$, having radius $r = 1$ (unit sphere). Every point $\boldsymbol{x}$ on the surface of the sphere is uniquely determined by a pair of angular coordinates: θ and λ where, by definition of spherical coordinates, we have: $(\theta, \lambda) \in \mathrm{I}^\ell_\angle = [-\pi; \pi] \times [-\pi/2; \pi/2]$. The notation $\mathrm{I}^\ell_\angle$ indicates the ℓ-dimensional, angular interval, in this case $\ell = 2$. The geodesic distance between any pair of points $\boldsymbol{x}_i, \boldsymbol{x}_k \in \mathrm{I}^\ell_\angle$ is given by (e.g. (Bomford 1980))

$$d_\Omega(\boldsymbol{x}_i, \boldsymbol{x}_k) = r\Delta\Omega \tag{3.7}$$

where r is the radius of the unit sphere: $r = 1$ and $\Delta\Omega$ is the central angle under the segment of great circle connecting $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$:

$$\Delta\Omega = 2 \arcsin \sqrt{\sin^2\left(\frac{\Delta\lambda}{2}\right) + \cos\lambda_i \cos\lambda_k \sin^2\left(\frac{\Delta\theta}{2}\right)}, \tag{3.8}$$

where $\Delta\lambda = \lambda_i - \lambda_k$ and $\Delta\theta = \theta_i - \theta_k$. In the spirit of the original GTM (Bishop et al. 1998b) we can now define a regular grid of size M on the angular interval $\mathrm{I}^\ell_\angle$, placing an RBF on each node $\boldsymbol{c}_m$, $m = 1, \ldots, M$ of the grid. The radial basis function centered on $\boldsymbol{c}_m$ is:

$$\phi(\boldsymbol{x}, \boldsymbol{c}_m) = \exp\left[\frac{-d_\Omega(\boldsymbol{x}, \boldsymbol{c}_m)^2}{2\sigma^2}\right] \quad , \tag{3.9}$$

where $\sigma > 0$ is a scale parameter. The only difference with (Bishop et al. 1998b) is the replacement of the Euclidean distance with the geodesic distance defined in Eq. (3.7) (hence, the name geodesic GTM). Every point on the unit sphere is then mapped to the ambient space by the function:

$$\boldsymbol{y}(\boldsymbol{x}; \mathbf{W}) = \mathbf{W}\boldsymbol{\phi}(\boldsymbol{x}), \tag{3.10}$$

where the elements of $\boldsymbol{\phi}(\boldsymbol{x})$ are the M RBFs as defined in Eq. (3.9) and $\mathbf{W}$ is the weight matrix of dimension $D \times M$.

The gGTM model will be trained on the "robust" set of boundary points, i.e. those that accumulated enough votes in the resampling voting scheme described in the previous section. We collect such points (exemplified in Figure 3.2b) in the set $\mathcal{Q} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_l\}$, where $v(\boldsymbol{t}_i) \geqslant v_3$ for all $i = 1, 2, \ldots, l$. To initialise the weight matrix $\mathbf{W}$ we first take advantage of a physics-inspired diffusion algorithm (SAF, (Wu et al. 2018b)) that collapses points $\boldsymbol{t_i} \in \mathcal{Q}$ toward high density regions in their proximity, thus sampling closer to the "mean" of the noisy manifold. The resulting data set $\tilde{\mathcal{Q}} = \{\tilde{\boldsymbol{t}}_1, \ldots, \tilde{\boldsymbol{t}}_N\}$ is the diffused version of the data $\mathcal{Q}$.

We then estimate the mean radius and the boundary centre as:

$$\overline{r_\mathcal{M}} = \frac{1}{2N}\sum_{i=1}^{N}\left[\max_k(\|\tilde{\boldsymbol{t}}_i - \tilde{\boldsymbol{t}}_k\|)\right]; \tag{3.11}$$

$$\overline{\boldsymbol{\mu}} = (\overline{\mu}^1, \overline{\mu}^2, \overline{\mu}^3)^\top = \frac{\sum_{i=1}^N \tilde{\boldsymbol{t}}_i}{N} \quad . \tag{3.12}$$

The weights in matrix $\mathbf{W}$ can be initialised by building a refined grid $\{(\theta_i, \lambda_i)\}_{i=1}^K$ over $\mathrm{I}^\ell_\angle$ and defining latent points $\boldsymbol{x}_i = (\theta_i, \lambda_i)$. The latent points are then mapped to the embedding space by applying the transformation between spherical and Cartesian coordinates:

$$\boldsymbol{\xi}_i = \begin{pmatrix} \xi_i^1 \\ \xi_i^2 \\ \xi_i^3 \end{pmatrix} = \begin{pmatrix} \overline{\mu}^1 + \overline{r_\mathcal{M}} \sin\theta_i \cos\lambda_i \\ \overline{\mu}^2 + \overline{r_\mathcal{M}} \sin\theta_i \sin\lambda_i \\ \overline{\mu}^3 + \overline{r_\mathcal{M}} \cos\theta_i \end{pmatrix} \quad . \tag{3.13}$$

The weights in matrix $\mathbf{W}$ are set through linear regression so that $\boldsymbol{y}(\boldsymbol{x}_i; \mathbf{W}) \approx \boldsymbol{\xi}_i$ for all corresponding points $\boldsymbol{x}_i \in \mathrm{I}^\ell_\angle$ and $\boldsymbol{\xi}_i$. For every point $\boldsymbol{x}_i$ in the latent space, the orthogonal vector to the spherical surface computed at the embedded point $\boldsymbol{y}(\boldsymbol{x}_i; \mathbf{W})$ is:

$$\hat{\boldsymbol{n}}_i = (\hat{n}_i^1, \hat{n}_i^2, \hat{n}_i^3)^\top = \frac{\boldsymbol{y}(\boldsymbol{x}_i; \mathbf{W}) - \overline{\boldsymbol{\mu}}}{\overline{r_\mathcal{M}}} \quad . \tag{3.14}$$

A pair of tangent vectors to $\mathcal{M}$ orthogonal to $\hat{\boldsymbol{n}}$ and spanning the tangent space $T_\mathcal{M}(\boldsymbol{\xi}_i)$ to $\mathcal{M}$ at the point $\boldsymbol{y}(\boldsymbol{x}_i; \mathbf{W})$ can be recovered by differentiating $\boldsymbol{\xi}_i$ (Eq. (3.13)) w.r.t. θ and λ. After normalization to unit length, we obtain

$$\boldsymbol{u}_i = \frac{\partial \boldsymbol{\xi}_i}{\partial \theta} = (\cos\theta_i \cos\lambda_i, \cos\theta_i \sin\lambda_i, -\sin\theta_i)^\top;$$
$$\boldsymbol{v}_i = \frac{\partial \boldsymbol{\xi}_i}{\partial \lambda} = (-\sin\lambda_i, \cos\lambda_i, 0)^\top \quad .$$

We can now define the noise model for our gGTM by constructing covariance matrices $\mathbf{C}_i$ of multivariate Gaussians centered at images $\boldsymbol{y}(\boldsymbol{x}_i; \mathbf{W})$ of the latent centers. In particular, we construct $\mathbf{C}_i$ proportional to the matrices having as Eigenvectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$:

$$\mathbf{C}_i = \frac{1}{\beta}\mathbf{I} + \eta(\boldsymbol{u}_i\boldsymbol{u}_i^\top + \boldsymbol{v}_i\boldsymbol{v}_i^\top) \ . \tag{3.15}$$

Here, $0 < \beta < 1$ is a regularization term, $\mathbf{I}$ the identity matrix and η a scaling factor proportional to the distance between neighbouring nodes of the embedded grid. The manifold aligned probabilistic model takes the form of a constrained mixture model:

$$p(\boldsymbol{t}; \mathbf{W}, \beta) = \frac{1}{K}\sum_{i=1}^{K} p(\boldsymbol{t}|\boldsymbol{x}_i, \mathbf{W}, \beta, \eta), \tag{3.16}$$

where each component $p(\boldsymbol{t}|\boldsymbol{x}_i, \mathbf{W}, \beta, \eta)$ is defined by:

$$p(\boldsymbol{t}|\boldsymbol{x}_i, \mathbf{W}, \beta, \eta) = \frac{1}{[(2\pi)^D|\mathbf{C}_i|]^{1/2}} \exp\left(-\frac{1}{2}\Delta\boldsymbol{t}^\top\mathbf{C}_i^{-1}\Delta\boldsymbol{t}\right)$$

and $\Delta\boldsymbol{t} = \boldsymbol{y}(\boldsymbol{x}_i; \mathbf{W}) - \boldsymbol{t} \in \mathbb{R}^3$. The parameters can be trained via the Estimation Maximization (EM) algorithm (Dempster et al. 1977), as described in (Bishop et al. 1998a) for a manifold-aligned noise model. In the case of circles, the gGTM model is constructed analogously. The only only difference is that the latent space has circular structure parametrized by one-dimensional angular interval $\mathrm{I}^{\ell}_{\angle} = [-\pi; \pi]$ with $\ell = 1$.

3.3 Experiments

In this section, we address the comparison between ASAP as a preprocessing step for building a simplicial complex and other state-of-the-art methods that were proposed to decrease the resources needed for computing the TDA. To compute the PH of point sets we mainly use GUDHI (The GUDHI Project 2020), which is faster and more memory efficient due to the structure of the simplex tree than many other toolboxes (Otter et al. 2017). In order to obtain the Rips filtration on datasets with larger number of points, we build the simplicial complex using Ripser (Bauer 2019). We first discuss controlled experiments with known ground truth, followed by the results of ASAP on real-world data from an astronomical galaxy particle simulation.

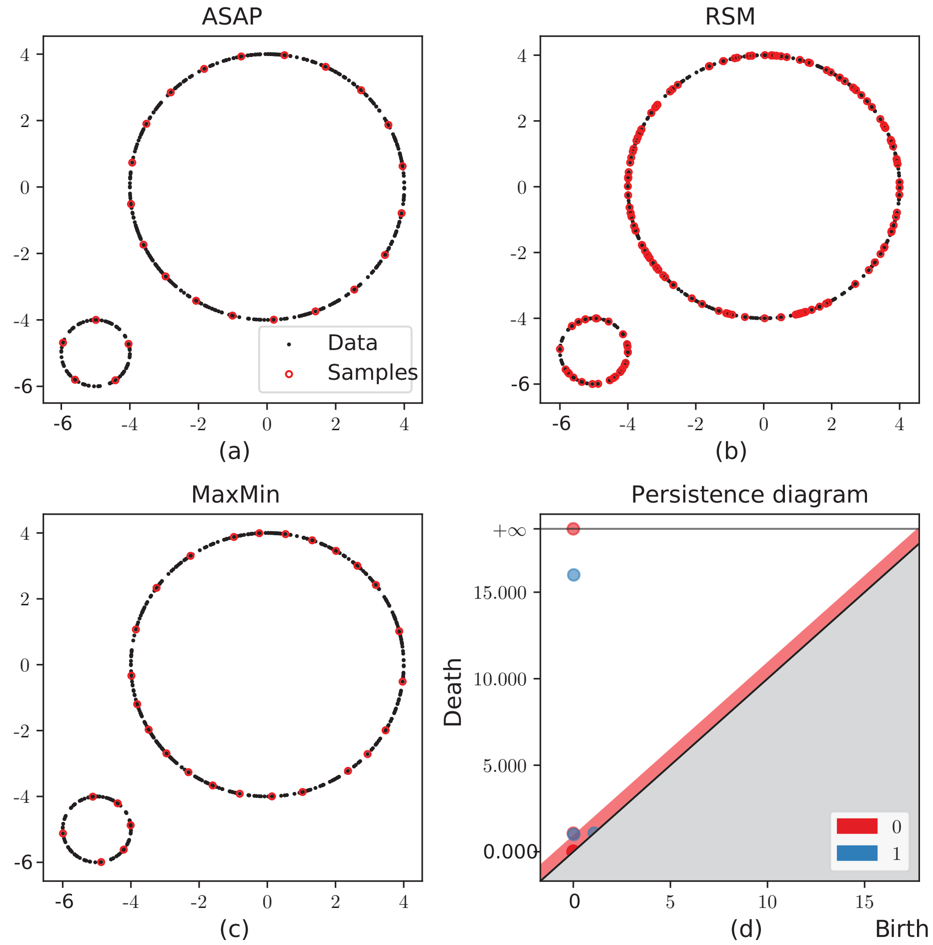

Figure 3.3: *2circles* dataset: The original 2D data (black) and subsampled data (red) as acquired by (a) ASAP (25 points), (b) RSM (Chazal et al. 2015) (175 points) and MaxMin (De Silva and Carlsson 2004) (30 points) that result in identical persistence diagrams as depicted in panel (d).

3.3.1 Synthetic data with ground truth

2circles dataset We first experiment on a simple two-dimensional dataset which was introduced in (Chazal et al. 2015) to demonstrate several sub-sampling methods for PH. 500 points are distributed uniformly on two circles with radius 1 and 4. For comparison we subsample 100 times with each method, saving the minimal point set required that recover the known features outside the 95% confidence band in the PH. We record the performance in form of several different evaluation measures, namely the average number of constructed simplices, as well as Median Relative Dominance (MRD) and Median Absolute Dominance (MAD), as summarized in table 3.1 and 3.2. Figure 3.3(d) shows the persistence plot for Alpha filtration of the samples selected based on ASAP, RSM (Chazal et al. 2015), and MaxMin(De Silva and Carlsson 2004). Each point illustrates the birth-death time of a topological feature of the point cloud. The points for features of homology groups H0, H1 are presented in red and blue, respectively. The death time of the features with Betti number 1 shows the correct value for the radii of both circles. The two red dots outside the confidence band manifest the data consists of two separated parts. Furthermore, the two blue dots outside the confidence band imply the existence of two significant holes in the dataset. The figure was denoised by removing points with minimum persistence (death time - birth time for every feature) smaller than threshold 0.5.

Notably, the sub-sampling suggested in (Chazal et al. 2015) can only reduce the number of samples to 175 points that preserve the persistence of all features of the original dataset, while MaxMin and the proposed method ASAP can reduce the original point set to only 30 and 25 points, respectively. However, comparing panel ASAP (a) and MaxMin (c) shows that the former samples are more uniformly spaced than MaxMin on both circles and therefore the PH over repeated runs is typically more robust. Furthermore, due to covering and packing conditions of ASAP, the distance between every two neighboring samples is not smaller than r and bigger than $2r$ if enough data points lie in the space between two samples. The same conditions does not hold for MaxMin samples.

In an additional experiment, we also compared the three sub-sampling methods reducing the number of points consecutively while observing the resulting death time of the small circle, with the result shown in Figure 3.4. Since we know the data points form a circle, the death time also stands for the radius of the circle. Note, that the circle with radius 4 contains more samples and hence is robust for all sub-sampling methods. We repeat the MaxMin and RSM 10 times as suggested in (Chazal et al. 2015), and the number of points in each sub-sample of ASAP corresponds to hyperparameter r ranging from $[0.1, 1.1]$. We are more lenient in this experiment, observing the death time without considering if the feature stands out of the 95% confidence band or not. For both Alpha and Rips filtration (Figure 3.4 (a) and (b)) ASAP and MaxMin preserve the death time of the smaller circle with much fewer points than RSM, as indicated by the shorter blue line. Moreover, in both plots by decreasing the number of points, the death time is mostly increasing.

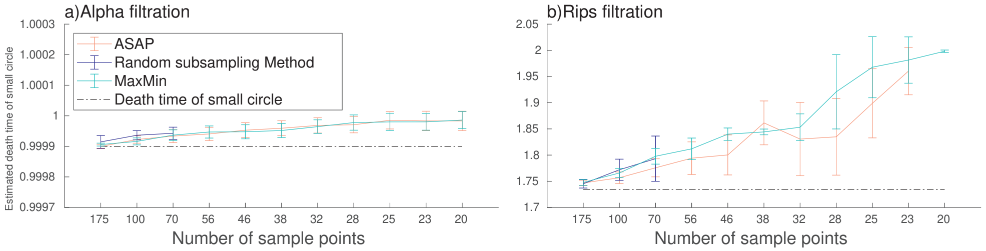

Figure 3.4: *2circles*: mean and standard deviation of the death time of the smaller circle over 10 repeated samplings of ASAP, RSM, and MaxMin when reducing the number of samples kept with Alpha (a) and Rips filtration (b).

In panel (a) the changes are not dominant, but in panel (b) the change is more visible with MaxMin diverging more from the baseline (depicting the true death time of the original small circle) for equal sample size.

2Spheres To compare the methods in higher dimensions we distribute points non-uniformly and unevenly on two hyper-spheres in $\mathbb{R}^5$ with radius 1 and 2, in the following referred to as *2Spheres* dataset. Even though the data consists only of 1200 points the computation of Rips filtration is very memory consuming due to dimensionality. Note that the code for efficient Rips filtration with SimBa (Dey, Shi and Wang 2019) only returns the Betti numbers up to 3 dimensions. We sub-sample the point cloud based on ASAP, RSM (Chazal et al. 2015), and MaxMin (De Silva and Carlsson 2004) and construct the Alpha complex on the resulting sub-sets. We iterate the sampling procedure 100 times and present the persistence diagram and barcode plot that is similar for all three methods in Figure 3.5. Since the data is not uniform RSM (Chazal et al. 2015) cannot preserve its homology outside the 95% confidence band if we reduce the number of samples to less than 1000 points. On the other hand, ASAP with radius 0.58 preserves not only the homology of the data in $\mathbb{R}^5$ with an average of 686 sub-sampled points, but also the death times for Betti number 4 corresponds to the radii of the spheres. MaxMin can also recover the same two features with a slightly lower number of samples 680. This is possible since MaxMin takes the number of samples as a parameter, while ASAP controls the number indirectly by the radius parameter. The persistence diagrams and barcode plots of all three methods are mostly identical if RSM is allowed enough samples, except for the small difference in the confidence interval, and therefore only the result of ASAP is displayed. The persistence barcode was furthermore denoised by removing properties with a minimum of birth-death time below 0.2 to make the plot more readable.

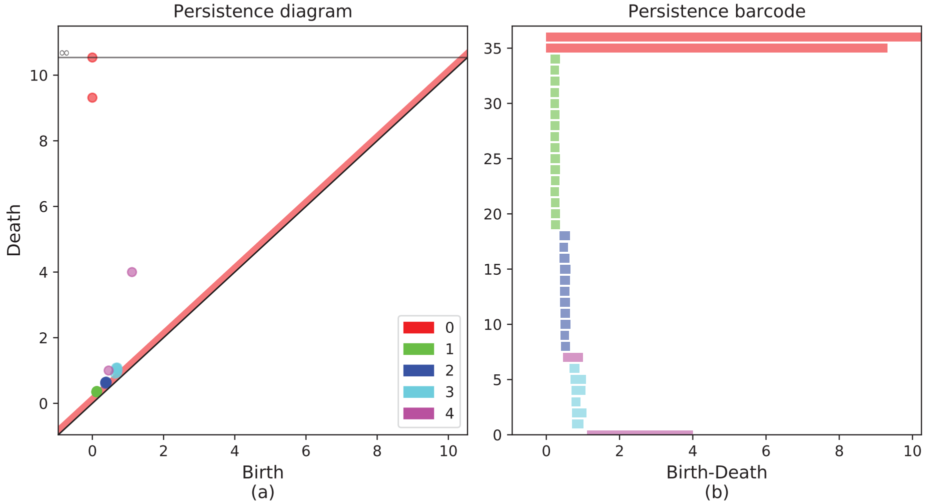

Figure 3.5: *2Spheres*: Persistence diagram (a) and barcode plot (b) of Alpha filtration. The plots are very similar for samples extracted by ASAP, RSM, and MaxMin and hence we just show one example.

Synthetic dwarf galaxy Finally we create a synthetic data-set mimicking some main characteristics of our astronomical application, in the following referred to as *s.dwarf*. In total 9656 non-uniformly distributed points in $\mathbb{R}^3$ form a synthetic dwarf galaxy containing 2 cavities with different size, 3 cycles: two with the same radius and one with a slightly bigger radius contained within a half spherical head, as well as a connected and a separated stream as shown in panel (a) of Figure 3.6. As demonstrated in panel (b), the persistence diagram of Alpha filtration on 551 sub-samples (only 5.7% of the original set) with ASAP ($r = 0.15$) preserves the main features of the original data and also maintains the radii of cycles and cavities. The RSM (Chazal et al. 2015), on the other hand, can save the same features outside the confidence band only with more than 6200 sub-samples.The striking difference between the number of samples needed using (Chazal et al. 2015) and ASAP is caused by the aim of the latter to distribute the points evenly and thus keeping the same topological features with much fewer samples. MaxMin can also preserve all known features outside the confidence band with 550 sub-samples. The difference between the persistence diagram of the three methods is small, thus we only present the ASAP result in panel (b).

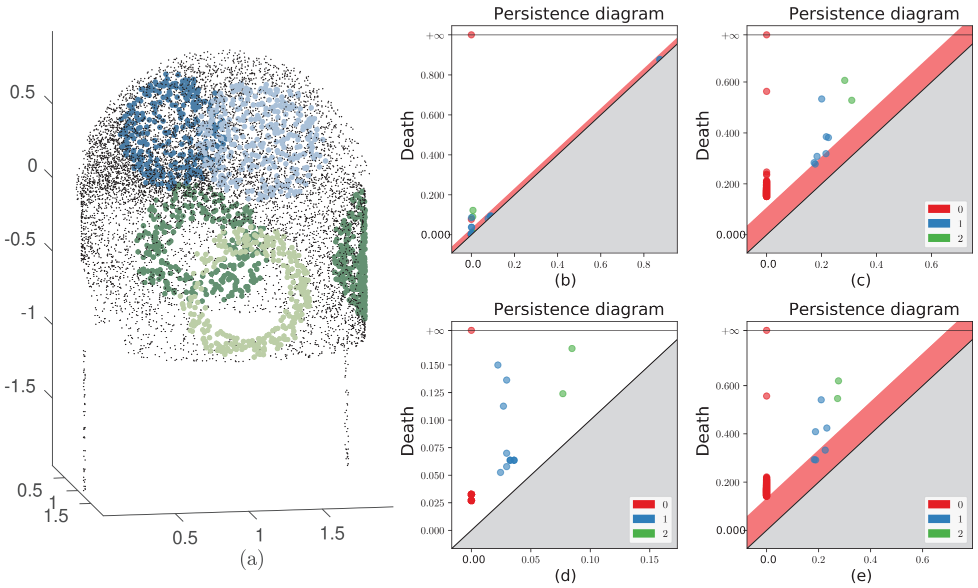

Figure 3.6: *s.dwarf* (a): persistence diagram of: (top) Alpha filtration based on 551 ASAP points (b), Rips filtration for ASAP points (c), (bottom) SimBa (d), and Rips filtration for 550 MaxMin points (e).

Both GUDHI and Ripser fail to compute the Rips filtration of the entire dataset due to high memory usage. Nevertheless, if we decrease the number of points by ASAP with $r = 0.15$, Ripser manages to calculate the Rips filtration and its persistence diagram. All expected features are visible outside the confidence band in panel (d). However, some new 0 homology features also appear in this plot that emerge since the distance between every two points after sub-sampling is more than r. SimBa is also capable of computing the Rips filtration, as depicted in Figure 3.6(d), but the death times do not conform with the exact size of the cycles and cavities within the head. Besides, 0 homology features are represented with three red points that do not correspond to the number of connected components. Note that points on the persistence diagram shape a multiset, and each red dot can illustrate more than one feature. We can also compute the Rips filtration on the MaxMin selected sub-samples. As presented in panel (e), all expected features stand out the confidence band, and similar to ASAP some extra 0 homology features are also included. Nevertheless, the birth time of features is more distorted as the features of the same size (two out of three cycles) are presented with two distinct points. The run time of the ASAP sampling for 2sphere dataset ($r = 0.58$) is about 0.4 second and it occupies about 1MB of memory. On the Synthetic dwarf galaxy ($r = 0.15$), it takes about 0.7 second to select the samples and it consumes 2.2MB of memory. All

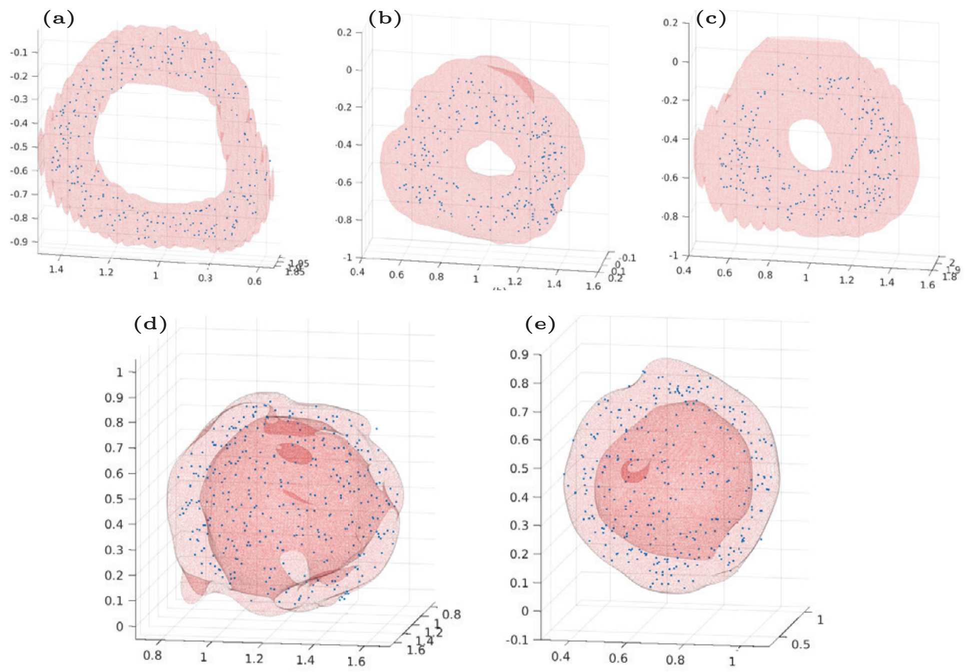

Figure 3.7: Iso-surfaces of the likelihood computed by the probabilistic models of the recovered cycles (top row) and holes (bottom) of the synthetic dwarf galaxy (s.dwarf) depicted in Figure 3.6(a).

experiments were conducted on a single core of a processor with a maximum clock rate of 4.5GHz.

To locate the position of notable features with a confidence value higher than 95%, we use Dionysus (*Dionysus, a C++ library for computing persistent homology* n.d.) and then applied the voting procedure as explained in section 3.2.3. In Figure 3.6(a) points in light and dark green are the detected points on the border of cycles, and the points in dark and light blue build the outline of two cavities inside the head of the synthetic galaxy. The probabilistic models visualized in Figure 3.7 built using the detected border points from Figure 3.6a clearly depict the intrinsic nature of the structures. The top panels (a-c) depict the likelihood of the datasets given the models, as iso-surfaces over the space containing the respective data points: the manifolds' neighbourhood. Each neighbourhood is discretized in a uniform grid and for each node in the grid we compute its likelihood given the manifold's model. The iso-surfaces in Figures 3.7-3.8 are obtained by locally interpolating all nodes of the grid having the same likelihood, equal to a specific iso-value. For all figures, the iso-value is chosen to be 1% of the maximum likelihood computed over the whole grid. The coherent structures emerging from these iso-surfaces explain the noisy cycles that characterize the boundaries of the 2-dimensional holes. In the same way, the noisy spherical iso-surfaces in panels (d and e) of Figure 3.7, cleanly separate regions populated by the boundary points (spherical shells) from the internal holes.

Table 3.1: Comparison of the number of simplices constructed by several methods and filtrations with lowest numbers marked in bold.

method \ dataset $(n\|d)$		2Circles (500|2)	2Spheres (1 200|5)	s.dwarf (9 659|3)
ASAP	Alpha	**81**	502 002	14 193
	RIPS	2 639	∞	-
RSM	Alpha	309	584 657	170 219
	RIPS	37 876	∞	-
MaxMin	Alpha	136	**496 123**	**14 125**
	RIPS	5 030	∞	-
SimBa	RIPS	**1 031**	-	63 004
GUDHI	Alpha	2 345	718 531	250 991
	RIPS	13 752 927	∞	∞

Table 3.1 presents the total number of simplices arising in every filtration on all synthetic datasets investigated. SimBa can only compute the Rips filtration and although Ripser computes the Rips filtration on the synthetic dwarf dataset it does not provide any information about the size of the simplicial complex inside the structure indicated by a "-" in the respective columns. Additionally, we denote with "∞" whenever the computation of the Rips filtration fails due to the memory complexity. For our proposed method, MaxMin (De Silva and Carlsson 2004) and RSM (Chazal et al. 2015), we report the results for the number of samples preserving the homology of the data after denoising. This information reveals that ASAP decreases the number of sample points significantly, and hence reducing the number of simplices in different filtration while preserving the topological features.

To evaluate the robustness of detecting known toplogical features from point clouds the Median Relative Dominance (MRD) and Median Absolute Dominance

(MAD) were introduced in (De Silva and Carlsson 2004). Relative dominance and absolute dominance are defined as $(R_1 - R_0)/K_0$ and $(R_1 - R_0)/K_1$, respectively, where R_0 and R_1 stand for the birth and death time of a feature. K_0 is the time when the Betti profile changes permanently to the profile of a single point in d-dimensions, and K_1 targets the time for which a complex becomes a complete simplex between all edges. Finally, we compute the median value over 100 iterations of sampling the data and calculating these metrics. Note that these metrics are calculated for every feature in a dataset separately, hence we add the suffix (s) and (b) in Table 3.2 to denote the results for the small and big circle or sphere respectively. The higher the value of these two metrics, the more robust the identification of similar features in the sub-sampled dataset is. Note that these metrics are not suitable for the synthetic dwarf galaxy dataset, since if the border points of the features are looser than the real border, the death time and metrics value increase falsely. Besides, the value of K_1 may vary drastically for alpha filtration if the center of an enclosing ball for the final added simplex located outside the simplex, thus we only discuss the MRD.

Table 3.2 displays the comparison based on these two metrics for ASAP and MaxMin. We did not insert RSM results here as long as RSM needed a larger number of samples to get similar topological features, thus the comparison is biased by the number of samples. As explained before, ASAP and MaxMin reach a similar number of samples for 2Spheres dataset. For 2Circles dataset, we chose the radius of sub-sampling using ASAP equal to 0.8 which results on average 28 samples that is closer to the number of samples selected by MaxMin (30). In both cases ASAP and MaxMin detect the expected features outside the 95% confidence band and the sub-sampling is repeated 100 times. The results disclose that ASAP reaches higher values for the MRD and MAD evaluation measures on both datasets. The lower metrics values for MaxMin stem from the strategy of the method to select samples. Figure 3.3 reveals that although MaxMin samples are more evenly spaced than RSM, they are not as well placed to their neighbors as ASAP samples, which lead to a later birth time and lower metrics values.

3.3.2 Particle simulation of a Jellyfish-like dwarf galaxy

Table 3.2: Comparison of MRD and MAD for several methods and filtrations with best values marked in bold. Suffix (s) and (b) mark the results for the small or big structure respectively.

	method / dataset \ metric / filtration	MRD Alpha	MRD RIPS	MAD RIPS
ASAP	2Circles(s)	**0.040**	**0.094**	**0.055**
	2Circles(b)	**0.973**	**0.813**	**0.474**
	2Spheres(s)	**0.110**	-	-
	2Spheres(b)	0.703	-	-
MaxMin	2Circles(s)	0.031	0.063	0.037
	2Circles(b)	0.962	0.780	0.459
	2Spheres(s)	0.097	-	-
	2Spheres(b)	0.703	-	-

Figure 3.8 panel (a) shows an N-body Smoothed Particle Hydrodynamics (Price 2012) simulation snapshot of a dwarf galaxy falling into a cluster environment with its gas stripped by ram pressure. The point set corresponds to the position of 33 500 gas particles. The distribution of points in this point cloud varies significantly, and points are dispersed on multiple separated parts. Hence, we expect to see several red points linked with Betti number 0 in the persistence diagram of this dataset as the Betti number 0 corresponds to connected components of the dataset. We sub-sample the dataset using ASAP with $r = 0.4$ reducing the set to $\approx 37.6\%$ of the total amount of points. Then the Alpha simplicial complex was constructed on the subset. We select a radius value for sub-sampling using ASAP to pursue two conditions: first, the expected topological features are outside the 95% confidence band and second, the computation of Alpha filtration on the remaining samples is tractable.

Figure 3.8 panel (b) shows the persistence diagram for the reduced set, denoised using a threshold of 1 for the minimum birth-death time. Based on the confidence band of 95% the data consists of four distinguished parts (0 homology features in red) and three cavities (blue points) with death time equal to 3.98, 1.66, and 1.48 respectively. We repeat the sub-sampling by ASAP 100 times, and each time, these three features are located inside the sub-sample sets, then using the technique defined in section 3.2.3, the points on the border of each hole are detected. Panel (a) also illustrates the border points of the three cavities inside the head part of the galaxy. The largest cavity has a late birth time (approximately 10) shown in panel (b) of Figure 3.8. A gap between points on the border of this cavity (points highlighted in light blue) is the reason for this delayed birth time. The first two holes were modelled via the modification to GTM described previously in sec. 3.2.4. The resulting iso-surfaces of the likelihood of the probabilistic models w.r.t. a regular grid for the points enclosing the two holes are shown in panel (d) and (e) of Figure 3.8. Given the spherical latent space adopted in our version of GTM, we could not properly model the irregular hole previously mentioned. Instead, we adopted

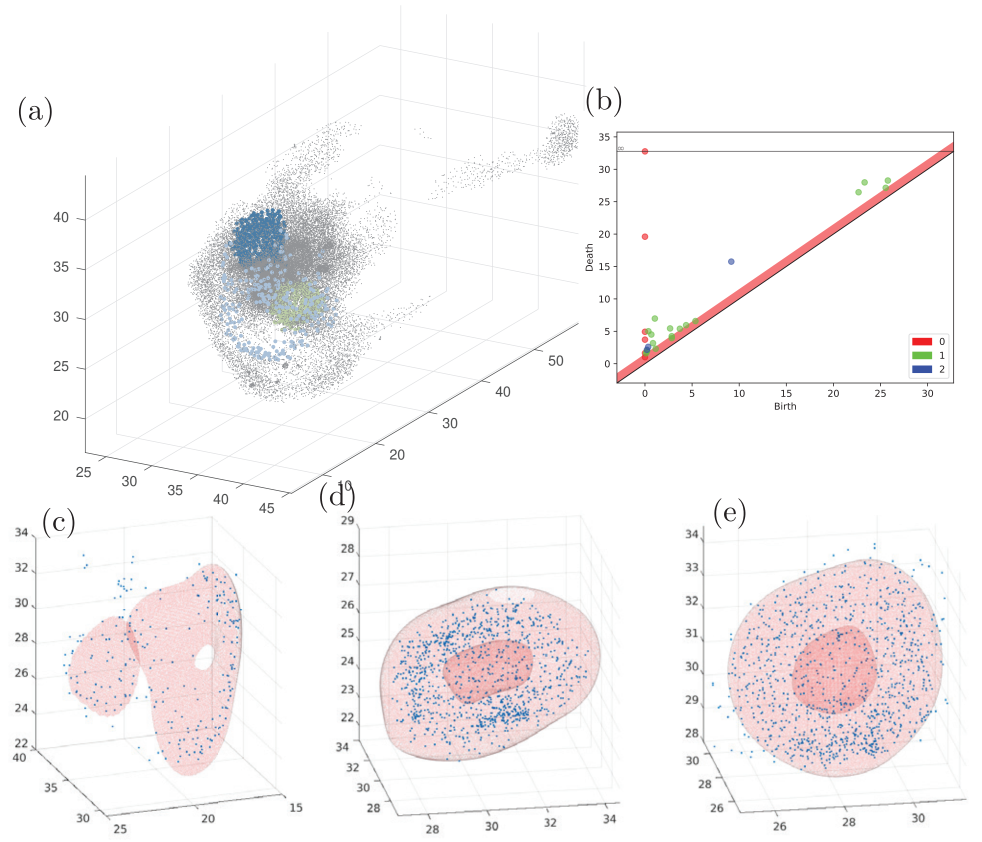

Figure 3.8: Jellyfish-like dwarf galaxy particles (a) and Alpha filtration of the ASAP subset (b). Iso-surfaces of the likelihood computed by the probabilistic models of the recovered holes (c-e).

the methodology outlined in (Wang et al. 2008), where multiple manifolds in a data set are modelled as low-dimensional graphs and embedded through the RBF formulation onto the ambient space. The results for this “hole” are shown in Figure 3.8(e).

The technique presented can be used to get insights into the physical processes at play in galaxy evolution by post-processing N-body simulations. Firstly, in the simulations, by computing the time evolution of the probabilistic iso-surfaces (like those shown in Figure 3.8) it is possible to follow the expansion of the gas around the modeled supernovae explosions, thus verifying the efficiency of the stellar feedback process. Also, distributions of the diameters, expansion velocities, ages of the

bubbles can be computed. The standard model describing the evolution of superbubbles is an adiabatic, pressure-driven expanding process with a continuous energy injection (Oey and Clarke 1997a). Recent simulations suggest that the ambient pressure does affect the expansion of the bubbles (Nath et al. 2020), which can now be analyzed quantitatively.

With ASAP and subsequent probabilistic modelling identifying the expanding superbubbles automatically, it would be possible to study the effect of the environment on the superbubbles accretion rate, and whether it depends on parameters such as the local ambient pressure and density. Moreover, the superbubble size distribution can be measured in a dynamical scenario like the one that is recreated in the simulations we used i.e. the fall of a galaxy in the hot and high density cluster gas. Observations show that the slope of the distribution of the size of superbubbles is different in galaxies with different morphologies (late-type vs. early-type) (Nath et al. 2020, Bagetakos et al. 2011). Our simulation scenario effectively captures the well known galactic morphological transformation due to the galaxy-cluster interaction, thus allowing a comparison of superbubble distribution for different galactic evolutionary stages. Lastly, being able to isolate the particles belonging to the cavity walls can shed light on the physical properties of the shock wave at the border of the bubbles.

3.4 Conclusion

In this chapter we expand the novel ASAP formulation for sub-sampling a point cloud that preserves the topological properties and reduces the memory consumption and computational cost for TDA analysis. The formulation is expandable for d-dimensions, is not limited to a specific type of filtration and its performance is shown for a variety of data sets. The features found are analyzed for their robustness using a statistical approach providing the confidence levels. We separate the signal from noisy features through a statistical test and argue the downside of the suggested technique for detecting the boundary of a cycle. Accordingly, we suggest a voting strategy to solve this problem, and finally, the points on the outline of the located cycles or cavities are modeled by a probabilistic model. Each model is generated by a generalized GTM approach, which allows further investigation and analysis of their properties. As it is disclosed through empirical results on several datasets, the proposed approach preserves the size of topological properties. The accuracy of such information is indispensable in some domains, such as astronomy where it informs about physical phenomena, namely supernovae in a galaxy.