Global and Relative Topological Features from Homological Invariants of Subsampled Datasets

Jens Agerberg [1] **Wojciech Chacholski** [1 2] **Ryan Ramanujam** [2 3]

Abstract

Homology-based invariants can be used to characterize the geometry of datasets and thereby gain some understanding of the processes generating those datasets. In this work we investigate how the geometry of a dataset changes when it is subsampled in various ways. In our framework the dataset serves as a reference object; we then consider different points in the ambient space and endow them with a geometry defined in relation to the reference object, for instance by subsampling the dataset proportionally to the distance between its elements and the point under consideration. We illustrate how this process can be used to extract rich geometrical information, allowing for example to classify points coming from different data distributions.

1. Objective

Topological data analysis (TDA) is about extracting topological signatures from samples in a dataset, for instance with the goal of feeding them as features to machine learning algorithms. For example, consider a dataset of greyscale images such as MNIST, which can be analysed as follows.

Via *individual approach*, where signatures are extracted for each image separately, for instance by regarding them as the grid of pixels that constitutes each image as a simplicial complex on which a real-valued function is defined based on the pixel intensity. Applying persistent homology to the super-level sets of this function, leads to features encoding some geometrical information about each image, such as the presence of a hole for a digit “0”.

Global approach focuses on the whole dataset, seen as a point cloud, with the aim to reveal global information about the distribution from which the dataset was drawn.

Relative approach, which we introduce in this article, focuses, as in the global approach, on the point cloud representing a dataset (or more generally, any reference object) but with a goal, similar to the individual approach, to reveal information about a particular sample. The sample defines a probability distribution on the dataset (e.g. by attaching higher probabilities to points near the sample), subsets of the dataset are then drawn according to the probability, from which topological signatures are extracted and averaged. The result is a description of the geometry of the sample, relative to the dataset, which can further be used to explore the geometrical variation of a dataset or be used in machine learning and statistical analysis.

This note is about illustrating a pipeline realizing the relative approach. We would like to stress that this note is not about how to choose parameters for the proposed pipeline. That will be the content of the follow up writings.

2. Introduction

Providing suitable representations of data by objects amenable for statistical and machine learning methods is one of the key steps towards a successful analysis. During the last decade homological representations of data played a significant role in many applications ranging from nano-characterization of materials (Lee et al., 2017) to neuroscience (Kanari et al., 2018). It is not well understood why homology-based invariants can be informative, and exploring this question has been difficult because homology is not directly amenable for statistical tools, as we can not average homology or describe what expected homology might mean. Thus for data analysis purposes homology needs to be transformed into objects that are amenable for statistical analysis. The aim of this article is to show that one such transformation, called *stable ranks* (introduced in (Scolamiero et al., 2017; Chachólski & Riihimäki, 2020)), can encode rich geometrical information about a dataset as a whole but also about a sample, relative to the dataset, when the latter is subsampled it in appropriate ways.

[1] Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden. [2] DatAnon, Corporation. [3] Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden.. Correspondence to: Jens Agerberg <jensag@kth.se>.

Proceedings of the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

datasets subsampled in various ways.

Consider a point in the plane. In itself a point does not have an interesting geometry, however in relation to other objects (called *reference* objects) it has rich geometrical aspects such as being on the left or right side of an oriented line, or being inside or outside a circle. Similarly, consider the very classical problem of recognizing handwritten digits in MNIST data (LeCun et al.). To decide if a handwritten digit represents 1 or 7, we might look at the geometrical aspects of the point representing the digit relative to, for instance, the reference object formed by some representatives of other handwritten digits. For both reading and writing purposes, humans have learned which variations of written digits can be recognized. Thus the key information needed to identify a handwritten digit is encoded in the spaces of points representing each digit. It is therefore expected that geometrical aspects relative to these spaces contain rich discriminative information. We propose to use homology-based invariants (called stable ranks) to encode this information and explore how they can be used to distinguish different types of points.

What do we mean by geometrical information? In this article it is information extracted from any space associated with the considered problem or object, where by a space we mean a simplicial complex or a family of them. For example let $\mathcal{R}$ be a finite subset of $\mathbb{R}^n$. By restricting a metric from $\mathbb{R}^n$ to $\mathcal{R}$ we can form its Vietoris-Rips complexes (Hausmann, 1995). This is just one instance of a possible spacial representation of metric properties of $\mathcal{R}$. Here is another possibility which we explore in this article. Choose a natural number s (called *sample size*) and consider the collection of the Vietoris-Rips complexes of some s-element subsets of $\mathcal{R}$, chosen according to some specified rule, for instance selecting all of them, or only those that are at certain distance to a given point. We find that this type of spacial representation of metric properties of $\mathcal{R}$ is informative. The way we use this representation is as follows. First, persistent homology is extracted for each of the subsets (there is an extensive literature regarding persistent homology see for example (Edelsbrunner & Harer, 2008; Ghrist, 2008; Weinberger, 2011)). If s is relatively small, the computational cost of this step is reasonable and can be done in parallel for each subset and for a rather high homological degree. The next step is to average these outcomes over some choice of the s-element subsets. This step requires transforming persistent homologies into objects whose averages (expected values) can be calculated. For that purpose we utilise stable ranks (Chachólski & Riihimäki, 2020; Gäfvert & Chachólski, 2017; Scolamiero et al., 2017) that represent persistence modules by non-increasing piecewise constant functions (see Appendix A). Stable ranks also enable us to use the associated kernel (Agerberg et al., 2021) and for example SVMs for classification purposes. We should mention that there are other ways of representing persistence modules by objects suitable for statistical analysis, for example persistence landscapes (Bubenik, 2015) or persistence images (Adams et al., 2017). The Vietoris-Rips persistent homology associated to a set of samples obtained by subsampling a point cloud has previously been explored in (Chazal et al., 2015; Solomon et al., 2022).

In this article a pipeline for assigning a stable rank (non-increasing piecewise constant function) to a subset $\mathcal{R}$ in $\mathbb{R}^n$ is presented. The space of parameters for our pipeline naturally splits into two types called *global* and *relative*. Global stable ranks encode some geometrical aspects of $\mathcal{R}$. Relative stable ranks encode some geometrical aspects of points in $\mathbb{R}^n$ relative to $\mathcal{R}$. In Section 5 we give simple examples in which it is possible to geometrically interpret the outcomes of the relative pipeline (lying on specified side of a hyperplane or inside a circle). Although in general geometrical interpretation of both relative and global stable ranks may be too complex, these invariants can be used for distinguishing purposes. For example we show that the training set in MNIST representing 7 is geometrically different from the test set in MNIST representing 7, as their invariants based on the homology in degree 1 are quite different. Thus, the geometry of the testing dataset for 7 is not entirely representative of the geometry of its training dataset. In Section 4 we also illustrate variability among global stable ranks of MNIST training datasets across different digits in homological degrees 0, 1, and 2. We explore this variability in Section 5 for classification purposes. For example if as a reference object we choose the union of the training MNIST datasets for digits 1 and 7, then the test digits labeled by 1 and 7 can be quite accurately classified using their relative stable ranks and only few labeled samples. We believe this is a consequence of the fact that global geometrical aspects of the training MNIST datasets for 1 and 7, as measured by their global stable ranks, are quite different.

3. Pipeline

The initial input is a finite subset $\mathcal{R} \subset \mathbb{R}^r$ called a *reference object*. We are going to explain how to assign to it various non-increasing piecewise constant functions. These functional representations of $\mathcal{R}$ can be then used as inputs for various analysis pipelines such as SVMs.

Step A: probabilities.

The objective is to obtain a function $\operatorname{prob}\colon \mathcal{R} \to \mathbb{R}$, called *probability*, with the following properties: all its values are non-negative, and their sum $\Sigma_{x\in\mathcal{R}}\operatorname{prob}(x)$ is either 1 or 0. For example we could take the uniform probability which is the constant function with value $1/|\mathcal{R}|$. Here is another construction, divided into two steps A1 and A2:

Step A1: filter function.

The objective is to obtain a function $\text{filter}\colon \mathcal{R} \to \mathbb{R}$ called a *filter*. In our particular construction the input consists of a point p in $\mathbb{R}^r$ and a vector field $\mathcal{V}\colon \mathcal{R} \to \mathbb{R}^r$ on $\mathcal{R}$. In this article the focus is on two types of a vector field: a constant vector field, and a vector field $\mathcal{V}_c\colon \mathcal{R} \to \mathbb{R}^r$ determined by a point c, called *center*, in $\mathbb{R}^r$ which assigns to x in $\mathcal{R}$ the vector $\mathcal{V}_c(x) := c - x$ from x to c. For example, we could take c to be the point p or the center of mass of $\mathcal{R}$.

In this step, the output is the function $\text{filter}\colon \mathcal{R} \to \mathbb{R}$ assigning to x the following value:

$$\begin{cases} 0 & \text{if } \mathcal{V}(x) = 0 \\ |\text{proj}_{\text{span}(\mathcal{V}(x))}(p - x)| & \text{if } \mathcal{V}(x) \cdot (p - x) \geq 0 \\ -|\text{proj}_{\text{span}(\mathcal{V}(x))}(p - x)| & \text{if } \mathcal{V}(x) \cdot (p - x) < 0 \end{cases}$$

For example, for the vector field $\mathcal{V}_p$, the associated filter function assigns to x in $\mathcal{R}$ the distance between x and p.

Step A2: distribution and probabilities.

In this step, we need to choose a non-negative function $\mathcal{D}\colon \mathbb{R} \to \mathbb{R}$ (called distribution), used to obtain the following probability function $\text{prob}\colon \mathcal{R} \to \mathbb{R}$ which is the outcome of this step. Let $S = \Sigma_{y \in \mathcal{R}} \mathcal{D}(\text{filter}(y))$.

$$\text{prob}(x) := \begin{cases} 0 & \text{if } S = 0 \\ \mathcal{D}(\text{filter}(x))/S & \text{if } S \neq 0 \end{cases}$$

Step B: averaged stable ranks.

The objective is to obtain a non-increasing piecewise constant function representing the reference object $\mathcal{R}$.

Step B1: sub-sampling.

The function $\text{prob}\colon \mathcal{R} \to \mathbb{R}$, obtained in step A, is used to sample the reference object $\mathcal{R}$. For this purpose two natural numbers s and n need to be chosen, called respectively *sample size* and *number of instances*. The outcome of this step is a set $\mathcal{S}$ described as follows:

- If $s > |x \in \mathcal{R} \mid \text{prob}(x) > 0|$, then the outcome $\mathcal{S}$ is the empty set.
- If $s \leq |x \in \mathcal{R} \mid \text{prob}(x) > 0|$, then the outcome $\mathcal{S}$ is of size n whose elements are subsets of $\mathcal{R}$ of size s. Each of these subsets is a random choice (with replacement) of s elements from $\mathcal{R}$ according to the probabilities specified by the function prob.

Step B2: stable ranks.

In this step the set $\mathcal{S}$ is converted into a *stable rank* function in the following way:

- Every element σ of the outcome $\mathcal{S}$ of step B1, which is a subset of the reference object $\mathcal{R}$, is converted into the following persistence module (homology of degree l with coefficients in $\mathbb{F}_2$ of the corresponding Vietoris-Rips complex, with respect to the Euclidean distance):

$$t \mapsto H_l(\text{VR}_t(\sigma), \mathbb{F}_2)$$

- For every σ in $\mathcal{S}$, the obtained persistence module is transformed into a non-increasing piecewise constant function given by its *stable rank* $\widehat{\text{rank}}(\sigma)$ with respect to the distance type contours D_f/T, associated with the density $f\colon [0, \infty) \to (0, \infty)$, and truncated at $T \in [0, \infty]$, see Definitions 5.4 and 5.6 (Chachólski & Riihimäki, 2020) (see also Appendix A). In this article we take f to be the standard density function given by the constant function 1.
- The final outcome of the entire pipeline, which depends on the *homological degree* l and *truncation* T defined in this step, is the average of all these stable ranks across all σ in $\mathcal{S}$:

$$\widehat{\text{rank}}_{\text{prob},s,n,l,T}\mathcal{R} := \left(\Sigma_{\sigma \in \mathcal{S}} \widehat{\text{rank}}(\sigma)\right)/n$$

4. Global stable ranks

The results of the pipeline described in Section 3, when the outcome of step A is given by the uniform probability function, are called *global stable ranks* of the reference object. These global stable ranks encode aspects of the geometry of the reference object captured by homologies of its s-element subspaces. In this section we illustrate examples of global stable ranks for the MNIST dataset (LeCun et al.). Recall that MNIST is a dataset of handwritten digits widely used in machine learning, composed of 60000 training samples and 10000 test samples. The samples are considered as points in $\mathbb{R}^{784}$, since the images have $28 \times 28 = 784$ pixels. For every d in $\{0, 1, \ldots, 9\}$, consider two reference objects $\text{Test}_d \subset \mathbb{R}^{784}$ and $\text{Train}_d \subset \mathbb{R}^{784}$ formed by these handwritten digits in respectively the test and the training sets of MNIST which are labeled by d.

As illustrated in Figure 1, the reference objects Test_d and Train_d, for $d = 2, 7, 8$ have noticeably different global stable ranks, indicating that there is some variation in the geometry between the training and test datasets. Since there is a probabilistic step in our pipeline, the whole process is repeated 10 times to demonstrate stability of the outcome.

A measure of geometric similarity between the reference objects can be obtained by considering distances between the obtained stable ranks, for example by using the L_1 distance. We compute the average stable ranks corresponding to the training and test set respectively and present the distance between them, for each digit, in Table 1. To further investigate whether the difference corresponds to a dataset shift or is due to random factors we pool the training and the test set together and perform random partitions. This is done 10 times for each digit, average stable ranks are then computed and the distance between the training and test sets resulting from these random partitions is compared to the distances obtained for the original training and test split. The results indicate that the difference between training and test sets is not due to random factors alone. Perhaps it results from the way the dataset was originally partitioned (partitioned by writers, but several samples belong to each writer hence potentially introducing a bias).

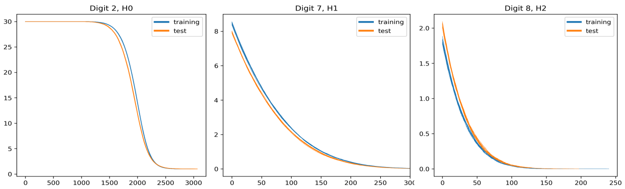

Figure 1. Stable ranks corresponding to a few digits, homology pairs, illustrating the difference between the stable ranks obtained from 10 repetitions of the pipeline for Test_d (orange) and for Train_d (blue) reference objects respectively. The following parameters were used:

prob	s	n	homological degree	T
uniform	30	2000	0 (left), 1 (middle), 2 (right)	∞

Table 1. Distances between stable ranks corresponding to training and test sets for different digits and homological degrees. The data is presented for stable ranks corresponding to the original training and test split and for stable ranks corresponding to random splits into training and test set.

Digit	H0 org	H0 rand	H1 org	H1 rand	H2 org	H2 rand
0	356.3	122.3	36.1	7.7	7.3	1.6
1	365.3	93.8	10.1	3.4	0.2	0.6
2	1143.7	125.7	30.1	6.7	4.6	0.6
3	982.7	134.3	15.0	3.6	1.1	0.8
4	728.0	41.6	27.2	1.7	6.1	0.4
5	294.7	119.8	54.1	13.5	7.9	3.3
6	550.1	42.2	31.0	2.4	2.6	0.5
7	702.6	101.0	49.5	2.3	4.9	0.3
8	427.5	132.6	46.5	9.4	7.5	0.8
9	843.2	57.1	24.3	4.8	2.2	0.8

When we write a digit, we intuitively know which variations still enable communication. We can think about the space Train_d as a space encoding such possible variations of d. A basic question is how dependent these spaces are on the digits and whether these spaces, for different digits, have detectable global geometrical differences. Figure 2 illustrates some global stable ranks of these spaces.

We note that in our experiments, the global stable ranks obtained by subsampling s-element subspaces were as (and sometimes more) distinctive of the digits as the stable ranks one can obtain from the computation of persistent homology on the whole reference object, without subsampling, a procedure that is heavier computationally.

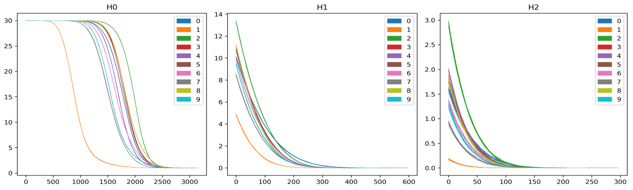

Figure 2. The pipeline was repeated 10 times for every reference objects Train_d, for all digits d, with the following parameters:

prob	s	n	homological degree	T
uniform	30	2000	0 (left), 1 (middle), 2 (right)	∞

In Section 6, we discuss a strategy of how to use the geometry of the spaces Train_d, encoded through our pipeline, to classify handwritten digits. Presented examples suggest that the further apart the geometrical properties of the spaces Train_{d_1} and Train_{d_2} are, the easier it is to distinguish between handwritten digits labeled by d_1 and d_2. This indicates that we need to look for ways of amplifying geometrical differences, if there are any, between the spaces Train_d for various d. In Figure 2 the stable ranks for some of the digits, such as digits 3 and 5, are hard to distinguish. Are these spaces then geometrically different and if so how can we encode differences between them? Let us change the truncation parameter T to 1800. The effect is shown in Figure 3, illustrating the fact that varying the parameters, e.g. sample size or the parameters used to construct the stable ranks, can lead to stable distinctive descriptors.

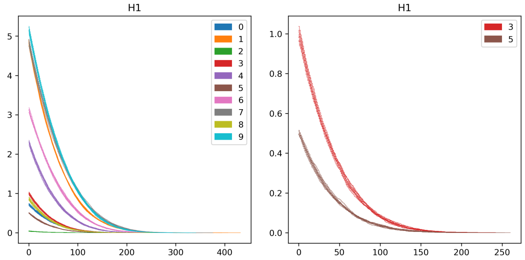

Figure 3. The pipeline was repeated 10 times for every reference object Train_d, for all digits d (left), and for digits 3 and 5 (right), with the following parameters:

prob	s	n	homological degree	T
uniform	30	2000	1	1800

5. Relative stable ranks in the plane

The results of the pipeline described in Section 3, when the outcome of step A is given by the probability function determined by a point p, are called *relative stable ranks* of the reference object. We think about relative stable ranks as encoding geometrical information about the position of the point p in the ambient space $\mathbb{R}^r$ in relation to the reference object. In this section we illustrate how relative stable ranks can be used to describe simple geometrical aspects of points in $\mathbb{R}^r$. Our initial data X consist of 200 random points on the plane (consisting of both orange and blue points in Figure 4) whose positions we would like to geometrically describe.

Example 1

- *Reference object*: a single point with coordinates $[-1, 2]$.
- *Point*: any point p in X.
- *Vector field*: given by the vector $[1, 1]$.
- *Distribution*: we consider two distributions:
  - $$\mathcal{D}_1(x) := \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$
  - $$\mathcal{D}_2(x) := \begin{cases} 1 & \text{if } -2 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
- *The other parameters*: $s = 1$, $n = 1$, homological degree is 0, and $T = \infty$.

Since the reference object consists of just one point, other homological degrees than 0 are irrelevant. The outcome of our pipeline in this case, for every point p in X, is a constant function 0 or 1. In this way the initial dataset X is partitioned into two clusters: points leading to the stable rank 0 and points leading to the stable rank 1. The two illustrations in Figure 4, which correspond to the two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, show such partitions of X. We see that our pipeline can for example distinguish between points lying on different sides of a hyperplane, an interesting piece of geometrical information.

Example 2

- *Reference object*: a noisy circle (of radius 3) represented by green dots in Figure 5.
- *Point*: any point p in X.
- *Filter*: assigns to an element in the reference object its distance to p.
- *Distribution*: Gaussian centered at 0 with standard deviation 1.

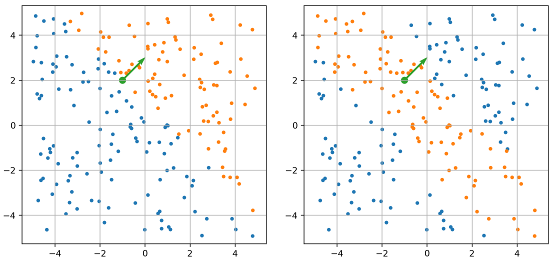

Figure 4. 200 random points colored according to whether the corresponding stable rank has constant value 1 (orange) or 0 (blue). The stable ranks were obtained with reference object containing only point $[-1, 2]$ and the vector field given by the vector $[1, 1]$ (Example 1 Section 5).

- *The other parameters*: $s = 10$, $n = 100$, homological degree is 0, and $T = \infty$.

In Figure 5 on the left, obtained stable ranks for all points in X are plotted. Those stable ranks corresponding to points whose distance to the origin is less than 3 are orange and the others are blue. In the illustration on the right a point is orange if corresponding stable rank at 0.87 has value bigger than 1.87. The other points of X are blue. Green dots represent the reference object. In this case a simple threshold obtained by visually inspecting the stable ranks allowed us to discriminate between points inside and outside the circle, again an interesting geometrical property. In the next section we will see that such classification rules can also be learned from the data.

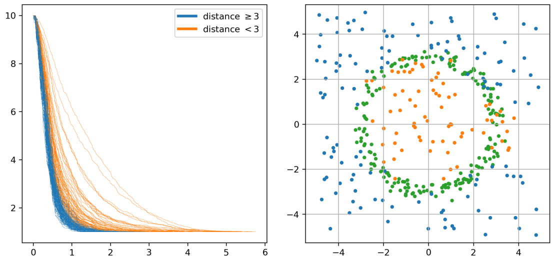

Figure 5. **Left**: Stable ranks corresponding to the random points in the plane, colored according to their distance to the origin. **Right**: Reference object (green) and random points colored according to whether the corresponding stable rank at 0.87 has value bigger (orange) or lower (blue) than 1.87 (Example 2 Section 5).

Deciding if a points is inside or outside a circle can be obtained by our pipeline with another set of parameters:

Example 3

- *Reference object*: the noisy circle as in Example 2.
- *Point*: any point p in X.

- *Vector field*: assigns to an element in the reference object the vector from that element to the center of mass of the reference object (which in this case is close to the origin).
- *Distribution*: $\mathcal{D}(x) := \begin{cases} 1 & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \end{cases}.$
- *The other parameters*: $s = 10$, $n = 100$, homological degree is 0, and $T = \infty$.

In Figure 6 on the left, obtained stable ranks for all points in X are plotted. As in Example 2, those stable ranks corresponding to points, whose distance to the origin is less than 3, are orange and the other are blue. In the illustration on the right a point is orange if the corresponding stable rank at 1.5 has value bigger than 3.9. The other points of X are blue, and the green dots represent the reference object. We see again that our pipeline can be used to decide if a point is inside or outside a circle.

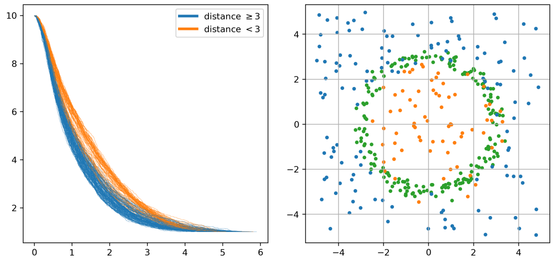

Figure 6. **Left**: Stable ranks corresponding to the random points in the plane, colored according to their distance to the origin. **Right**: Reference object (green) and random points colored according to whether the corresponding stable rank at 1.5 has value bigger (orange) or lower (blue) than 3.97 (Example 3 Section 5).

6. Relative stable ranks on MNIST

In Section 4 we provided experiments on MNIST for global stable ranks. In this section we now shift the focus to relative stable ranks. In the following experiments, subsets of the training sets corresponding to one or several digits will be used as reference objects. We will illustrate that these reference objects, when sampled from the perspective of different types of points in $\mathbb{R}^{784}$, such as points corresponding to digits in the test set, have interesting geometries.

Following the steps defined in the pipeline, for a point under consideration p and for elements in the reference object x in $\mathcal{R}$ we choose as *filter function* $f_p(x) = ||p - x||_2$, i.e. the Euclidean distance between the point under consideration and the elements of the reference object. As *distribution* we choose a Gaussian, whose parameters μ_p, σ_p are chosen in order to concentrate the probability mass on elements of the reference object close to p, yet ensuring the probability mass is distributed on sufficiently many elements for the samples to be diverse enough. We consider the set Dist_p of all distances between p and points in $\mathcal{R}$, i.e. $f_p(x)$ for all $x \in \mathcal{R}$. We select μ_p to be the k:th percentile of Dist_p, where k typically is a low number. We then choose σ_p in relation to the *sample size* parameter such that *sample size* $\times$ *amplification* elements of Dist_p lie within one standard deviation, where *amplification* is also a fixed parameter.

6.1. Illustration of the pipeline and first example

We start with a basic example to illustrate the pipeline. We take as our reference object the set Train_1, of all samples from the MNIST training set corresponding to the digit 1. Next we select two points from the ambient space, $\mathbb{R}^{784}$: the origin of that space and the center of mass of the reference object. Based on the values of the filter function, a Gaussian is chosen for each of the two points (we use as parameters *k:th percentile*$= 1$, *amplification*$= 5$). Next, as described in Section 3, a probability distribution on the reference object is computed for each point, by evaluating the values of the filter function under the Gaussian and normalizing.

We illustrate this idea in Figure 7. The two first principal components of the reference object are computed. We then project the reference object together with the origin and center of mass on the principal components. The origin (left plot) and the center of mass (right plot) are represented by black squares, and the dots representing elements of the reference object are colored according to their probability in the same way as in the previous plot.

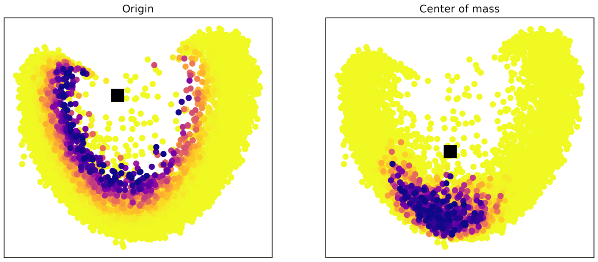

Figure 7. Projection on the two first principal components of the reference object. The origin (left) and the center of mass (right) are represented by black squares. Other dots are colored according to their probability.

Having illustrated that different points lead to different probability distributions, we now subsample the reference object according to these probability distributions. For each such sample a distance space is constructed (with Euclidean distance). Next, as described in Section 3, these distance spaces are converted into persistence modules corresponding to each homological degree, and then to stable ranks (we use *sample size*$= 50$, *number of instances*$= 100$). The resulting average stable ranks, presented in Figure 8, demonstrate that the geometrical signatures corresponding to the origin and the center of mass are distinct. We plot 10 stable ranks for each point and homological degree, obtained by repeating the whole procedure each time.

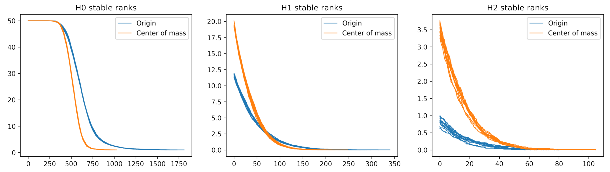

Figure 8. Stable ranks corresponding to the origin and the center of mass for different homological degrees.

Another way to illustrate how the geometry changes as we subsample the reference object in different fashions is to take the perspective of one point only – here we choose the center of mass – but to vary the parameters defined in the previous section. In Figure 9 we show the effect of increasing the *amplification* parameter, which means that less probability mass will be concentrated on elements whose distance is close to the mean of the Gaussian. As the value of this parameter increases, the stable ranks become closer to the global descriptor of the reference object described in Section 4, i.e. to the stable rank obtained by uniform subsampling of the reference object, indicating that the geometry is less and less informative.

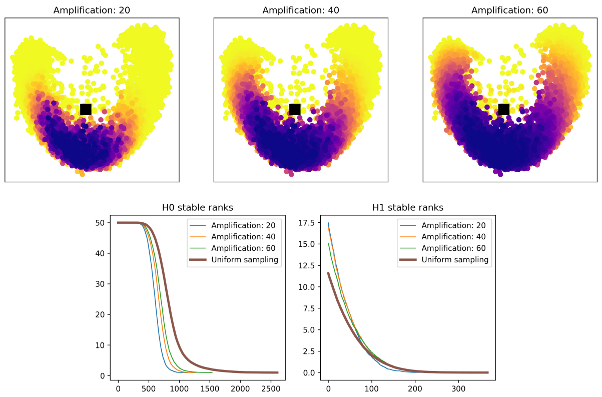

Figure 9. **Plot 1, 2, 3**: Projection on the two first principal components of the reference object. The center of mass is represented by a black square. Other dots are colored according to their probability when varying the *amplification* parameter. **Plot 4, 5**: Stable ranks corresponding to the center of mass for the different parameters and stable rank corresponding to uniform subsampling, for homological degrees H0 and H1.

6.2. Inside and outside

Instead of choosing the center of mass of the whole reference object, we now perform k-means clustering (k=10) on the reference object and select the center of mass of each cluster. We also sample 10 points randomly from the ambient space (the subset of $\mathbb{R}^{784}$ corresponding to allowed pixel values). We can then apply the procedure described in the previous section to obtain 10 stable ranks for the points corresponding to the centers of mass and 10 stable ranks corresponding to the random points, for each homology degree.

These stable ranks are displayed in Figure 10 together with the average stable rank corresponding to a uniform subsampling of the reference object. Our aim is to illustrate that stable ranks resulting from sampling from "inside" the reference object, e.g. for centers of mass, are distinct from stable ranks obtained by sampling from the "outside", e.g. from random points in the ambient space or from the origin (in the previous example). The latter are in turn more similar to the stable rank obtained by uniform subsampling.

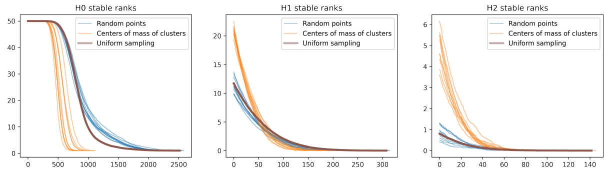

Figure 10. Stable ranks corresponding to random points, to the centers of mass of clusters and to uniform sampling, for different homological degrees.

6.3. Distinguishing out-of-sample points from two subsets of the reference object

In the previous section, we saw a clear distinction between stable ranks obtained by sampling from the "inside" and from the "outside" of the reference object. But the stable ranks corresponding to different centers of mass also displayed some variability, indicating a difference in the geometry. We aim to explore this idea further in the following setting: we now take as our reference object the union $\text{Train}_1 \cup \text{Train}_7$ of all samples from the MNIST training set corresponding to digit 1 or digit 7. We note that we still have only one reference object and the labels (indicating whether an element of the reference object corresponds to a 1 or a 7) are not used. Instead of considering random points or centers of mass as in the previous section, we now consider 10 points randomly chosen from Test_1 and 10 points randomly chosen from Test_7 and repeat the same procedure to compute the stable ranks representing these points. In our pipeline we use the following parameters: *sample size*$= 30$, *amplification*$= 2$, for the *homological degree* 0, the *trunca-*

tion parameter T is set to be ∞, and for the *homological degree* 1, the *truncation* parameter is set to be 1200.

We can see in Figure 11 that the stable ranks corresponding to test set digit 1 are distinct from those corresponding to test set digit 7, and they are both distinct from the stable ranks resulting from the uniform subsampling of the reference object. Hence, when we sample based on distances to test set 1 digits or 7 digits, we sample subsets of the reference object where the geometry is different, which allows us to discriminate between the points we sampled from.

To quantify the capacity to discrimate between digits based on their stable ranks, we train a Support-vector machine classifier on the 20 stable ranks, for each homological degree, using the kernel obtained by taking inner products between stable ranks in the L_2 function space (Agerberg et al., 2021). We can then evaluate the model on the remaining samples of digit 1 and 7 from the MNIST test set (samples that are neither part of the reference object nor part of the 20 samples used for the training). We obtain an accuracy of 96.9% for H0 stable ranks and 94.5% for H1 (average accuracy after repeating the procedure 10 times with different samples used for training). While we are not aiming at approaching state of the art accuracy levels we believe the results point to the fact that the geometry of a reference object, when chosen judiciously and in relation to a point, can be informative about characteristics of this point. We also note that we used a large unlabeled dataset (our reference object) but only a few (20) labeled samples, which is the setting of semi-supervised learning.

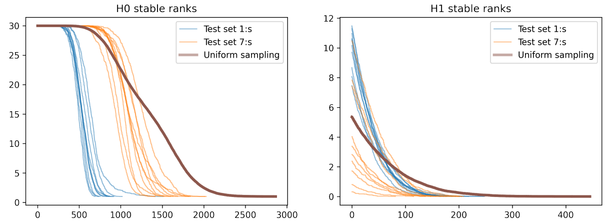

Figure 11. Stable ranks corresponding to test set 1 digits, 7 digits and to uniform sampling, for different homological degrees. Training set 1 digits and 7 digits used as reference object.

6.4. Distinguishing out-of-sample points based on another reference object

In the previous section, we considered a reference object which consisted of samples from the same data distributions (handwritten 1 digits and 7 digits) as the points that we sampled from and tried to discriminate. Now, while still trying to distinguish between test set samples of digits 1 and 7, we instead take as our reference object the union $\text{Train}_2 \cup \text{Train}_3$ of all samples from the MNIST training set corresponding to digits 2 or 3. Stable ranks are computed following the same procedure, however, for the *homological degree* 1, we used 1900 for the *truncation* parameter T. The obtained stable ranks are illustrated in Figure 12. Interestingly, when subsampled from different points representing 1 digits and 7 digits, the geometry of this reference object, which a priori is not related to the data distribution of those digits, nonetheless contains information about those points.

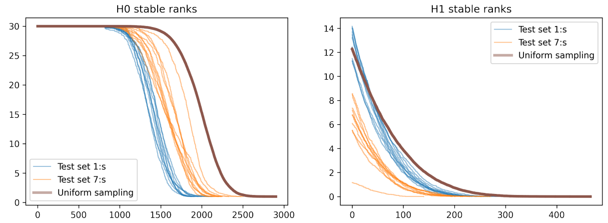

Figure 12. Stable ranks corresponding to test set 1 digits, 7 digits and to uniform sampling, for different homological degrees. Training set 2 digits and 3 digits used as reference object.

7. Discussion

Extracting stable ranks is a simplifying procedure. Finding appropriate parameters controlling stable ranks so that relevant aspects of the problem at hand are retained is the key challenge. In this paper we indicate that choosing an appropriate reference object and ways of sampling it can be used for this purpose. For example in experiment 6.3 the reference object is the union of the training sets corresponding to digits 1 and 7 which was shown to be effective for distinguishing between them. While analogous experiments can be repeated with similar results for several other pairs of digits, some pairs of digits were nonetheless harder to distinguish. In experiment 6.3 we could in general see that it was harder to distinguish test set digits from the two classes when the global geometries of the digits (see Section 4) were similar. But while more difficult, it was still often possible, since by sampling from the perspective of different points one can reveal different local geometric patterns that are specific to the digit. A classifier, when fed with such patterns, can thus still learn to distinguish the digits. Moreover, in experiment 6.4, when sampling a reference object that is not the union of the training sets corresponding to the digits we want to distinguish, we are in a different situation where global geometric similarity of the digits does not necessarily matter. Which reference object to choose is however not obvious. Another possibility is to combine different geometric signatures, e.g. stable ranks obtained by taking the training sets corresponding to digits 1 and 7 as separate reference objects, and computed for different homological degrees and parameters. Such signatures could then be combined in e.g. an ensemble learning scheme. We also emphasize that our method by construction only considers relative geometrical aspects to a point. Another interesting direction is thus to combine it with other methods (distance-based machine learning methods, neural networks, etc.) and analyze the combined effect.

Acknowledgments

JA was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation. WC was partially supported by the Wallenberg AI, Autonomous System and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation, and MultipleMS funded by the European Union under the Horizon 2020 program, grant agreement 733161, and dBRAIN collaborative project at digital futures at KTH. Support of the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation was indispensable for conducting this research. We would like to thank the reviewers for comments that helped improve the article.

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. Persistence images: a stable vector representation of persistent homology. *J. Mach. Learn. Res.*, 18:Paper No. 8, 35, 2017. ISSN 1532-4435.

Agerberg, J., Ramanujam, R., Scolamiero, M., and Chachólski, W. Supervised learning using homology stable rank kernels. *Frontiers in Applied Mathematics and Statistics*, 7:39, 2021. ISSN 2297-4687. doi: 10.3389/fams.2021.668046.

Bubenik, P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16:77–102, 2015. ISSN 1532-4435.

Chachólski, W. and Riihimäki, H. Metrics and stabilization in one parameter persistence. *SIAM J. Appl. Algebra Geom.*, 4(1):69–98, 2020. doi: 10.1137/19M1243932.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Subsampling methods for persistent homology. In *International Conference on Machine Learning*, pp. 2143–2151. PMLR, 2015.

Edelsbrunner, H. and Harer, J. Persistent homology—a survey. In *Surveys on discrete and computational geometry*, volume 453 of *Contemp. Math.*, pp. 257–282. Amer. Math. Soc., Providence, RI, 2008. doi: 10.1090/conm/453/08802.

Ghrist, R. Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc. (N.S.)*, 45(1):61–75, 2008. ISSN 0273-0979. doi: 10.1090/S0273-0979-07-01191-3.

Gäfvert, O. and Chachólski, W. Stable invariants for multidimensional persistence. *arXiv:1703.03632*, 2017.

Hausmann, J.-C. On the Vietoris-Rips complexes and a cohomology theory for metric spaces. In *Prospects in topology (Princeton, NJ, 1994)*, volume 138 of *Ann. of Math. Stud.*, pp. 175–188. Princeton Univ. Press, Princeton, NJ, 1995.

Kanari, L., Dlotko, P., Scolamiero, M., Levi, R., Shillcock, J., Hess, K., and Markram, H. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16:3–13, 2018. doi: 10.1007/s12021-017-9341-1.

LeCun, Y., Cortes, C., and Burges, C. J. *The MNIST database of handwritten digits*. URL http://yann.lecun.com/exdb/mnist/.

Lee, Y., Barthel, S., Dlotko, P., Moosavi, S., Hess, K., and Smit, B. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8 (15396), 2017. doi: 10.1038/ncomms15396.

Scolamiero, M., Chachólski, W., Lundman, A., Ramanujam, R., and Öberg, S. Multidimensional persistence and noise. *Found. Comput. Math.*, 17(6):1367–1406, 2017. ISSN 1615-3375. doi: 10.1007/s10208-016-9323-y.

Solomon, E., Wagner, A., and Bendich, P. From Geometry to Topology: Inverse Theorems for Distributed Persistence. In Goaoc, X. and Kerber, M. (eds.), *38th International Symposium on Computational Geometry (SoCG 2022)*, volume 224 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 61:1–61:16, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-227-3. doi: 10.4230/LIPIcs.SoCG.2022.61.

Weinberger, S. What is. . .persistent homology? *Notices Amer. Math. Soc.*, 58(1):36–39, 2011. ISSN 0002-9920.

A. Appendix

In this appendix we briefly recall the role the parameters *density* and *truncation* of our pipeline (see Section 3) play for constructing stable ranks. We refer to (Agerberg et al., 2021; Chachólski & Riihimäki, 2020; Gäfvert & Chachólski, 2017) where more information about stable ranks can be found.

Stable ranks are built using a process called hierarchical stabilization. An input for this process has two ingredients. One is a discrete invariant such as the rank function $\text{rank}\colon \text{Tame}([0,\infty), \text{vect}_{\mathbb{F}}) \to \mathbb{N}$, which assigns to a persistence module its minimal number of generators. The other ingredient is a pseudometric d on the domain of the discrete invariant, which in the case of the rank function is given by persistence modules $\text{Tame}([0,\infty), \text{vect}_{\mathbb{F}})$. The outcome of the hierarchical stabilization, for the mentioned rank function, is a Lipschitz function $\widehat{\text{rank}}_d\colon \text{Tame}([0,\infty), \text{vect}_{\mathbb{F}}) \to \mathcal{M}$, called *stable rank*, where $\mathcal{M}$ is the space of Lebesgue measurable functions $[0,\infty) \to [0,\infty)$. We think about the stable rank function as the model associated to the pseudometric d. In this framework (supervised) persistence analysis is about identifying these pseudometrics d for which structural properties of the (training) data are reflected by the geometry of its image in $\mathcal{M}$ through the function $\widehat{\text{rank}}_d$.

The reason we care about densities and truncations is because any choice of them leads to a pseudometric on persistence modules. Thus we can use densities and truncations as parameters of a rich space of such pseudometrics. We refer the reader to the mentioned sources for an explanation of how a density and a truncation leads to a pseudometric. See (Agerberg et al., 2021; Chachólski & Riihimäki, 2020) for examples where choosing an appropriate density leads to improvement in certain classifications tasks. In this article we have seen that a choice of truncation can also lead to better results.