Chapter 10

Applications

In this chapter we highlight some classical and recent applications of topological data analysis. The fields of application are as diverse as image analysis, medicine, analysis of time series and politics. Many of these applications are also described in the book "Topological Data Analysis with Applications" by Gunnar Carlsson and Mikael Vejdemo-Johansson [2]. We also discuss some of the approaches to use persistence diagrams in machine learning.

10.1 The Space of Image Patches

We consider a dataset of about 4000 greyscale images taken around Groningen (Netherlands) by van Hateren and van der Schaaf [8], see Figure 10.1 for some examples. Each such image is described by giving each pixel a number between 0 (white) and 1 (black). From every image, they sampled 5000 patches of 3-by-3 pixels. To get interesting patches, they only looked at the top 20% with the highest contrast. Thus, we end up with about 4 million three by three patches, each describable as a vector in $\mathbb{R}^9$. They further normalize the contrast (such that the darkest and brightest pixel are 1 and 0, respectively) and the overall norm of the vectors. Thus, we end up with points on S^7. How does this point cloud look? Is there some interesting structure in this point cloud?

Carlsson, Ishkhanov, de Silva and Zomorodian analyzed this point cloud using persistent homology [1]. Since there are a lot of points, they at first only sampled from the densest parts of the data. Computing the persistent homology of the Witness complex in dimensions 0, 1, and 2, one gets the barcodes depicted in Figure 10.2. It is thus a reasonable assumption that $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 0$, suggesting that the data set looks like a circle. PCA also shows a circular arrangement of the data points, see Figure 10.3. This circle can be interpreted as the possible angles of a detected "edge" in the images, depicted in Figure 10.4.

What happens if we additionally also sample some points from parts of the data of intermediate density? Computing the persistent homology we end up with the barcodes in Figure 10.5, suggesting that instead of one 1-cycle, we end up with five. If we again

Figure 10.1: *Some examples of images in the data set.*

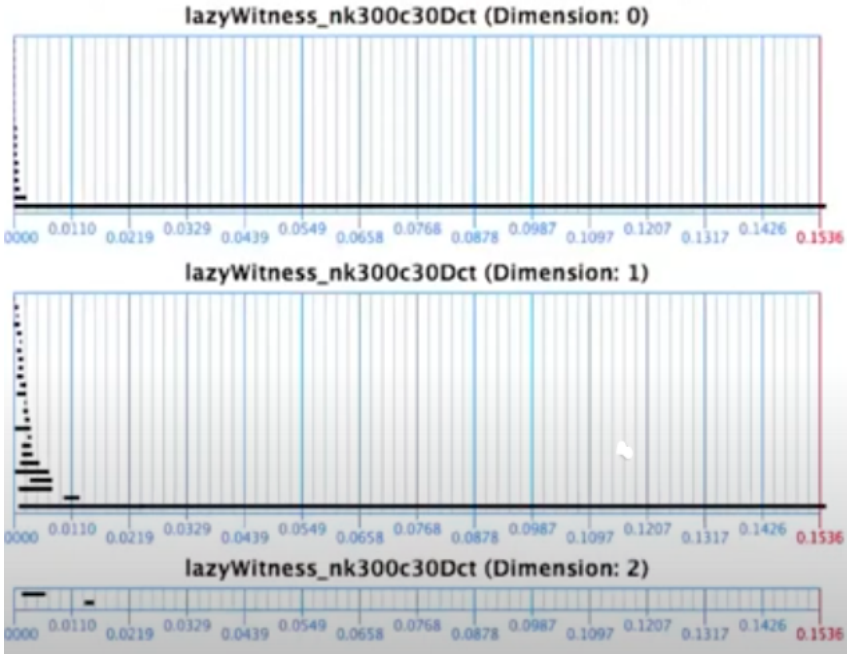

Figure 10.2: *The barcodes of the densest part of the data.*

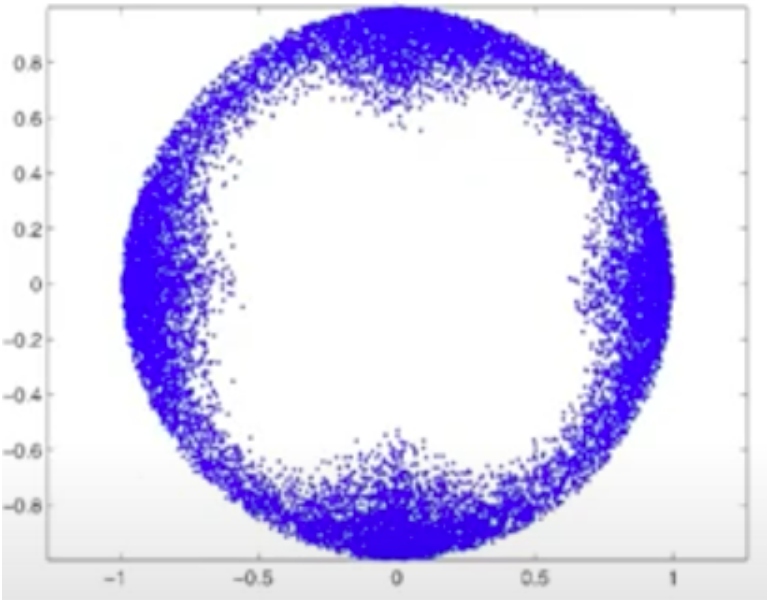

Figure 10.3: *Using PCA on the densest part of the data.*

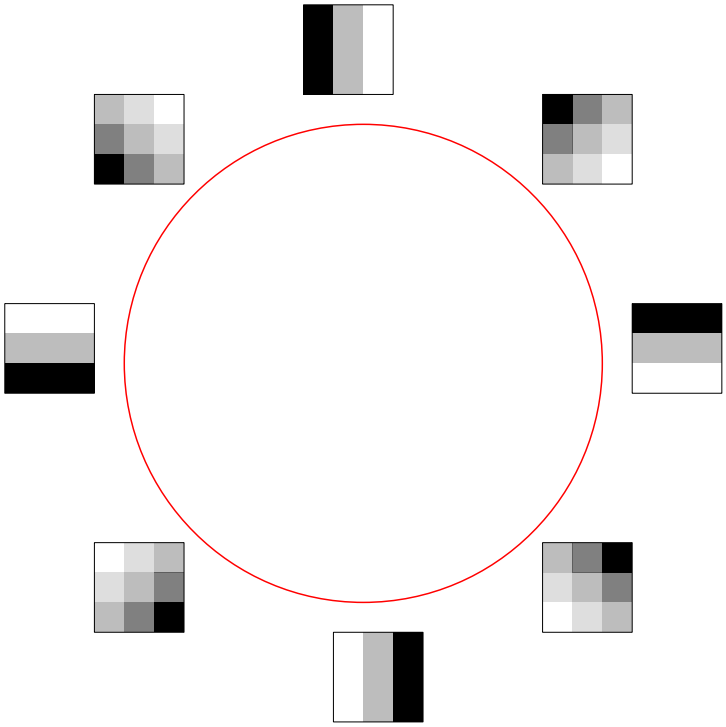

Figure 10.4: *The interpretation of the densest part of the data.*

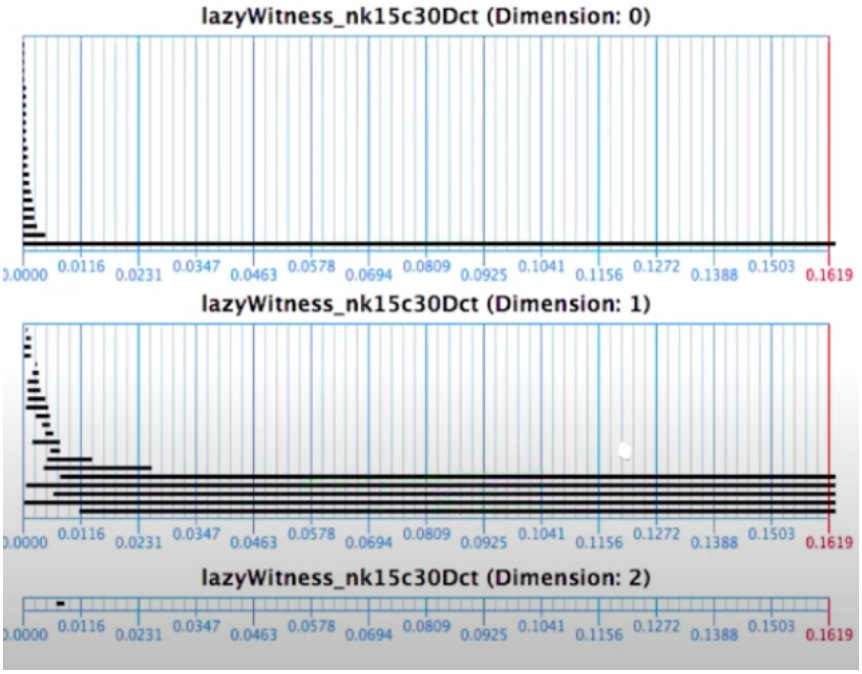

Figure 10.5: *The barcodes including the medium density part of the data.*

consider the PCA, depicted in Figure 10.6, we see a cross in the middle of the circle seen previously. However, this only shows four circles! Where is the fifth one? Most likely some cycle is orthogonal to the projection chosen by the PCA. The interpretation picked from this situation is that the dataset consists of three great circles of some S^2 that are all orthogonal. We can interpret these three circles as the circle described above, a circle describing different translations of vertical edges, and one circle describing translations of horizontal edges. This is depicted in Figure 10.7.

However, if we finally also include points from the lowest density parts of the data, we get the barcodes depicted in Figure 10.8. We start seeing some feature in the persistent 2-homology, and reduce β_1 to most likely 2 again. How can this space look? It must somehow include the three circles found before. We can embed the circles as in the Figure 10.9, indicating that the space is a Klein bottle.

Clearly, this is not a proof. The persistent homology computed agrees with the Klein bottle, but it would also agree with a torus. To give more evidence, we could for example compute persistent homology over a different field such as $\mathbb{Z}_3$, with which we can distinguish between a Klein bottle and a torus. It turns out that the space is indeed a Klein bottle, as is shown in the work of Carlsson et al. [1].

10.2 Mapper for Medical Data

The following application comes from the original paper describing the Mapper approach [6]. The data comes from a diabetes study in 1979 on 145 participants, with six quantities

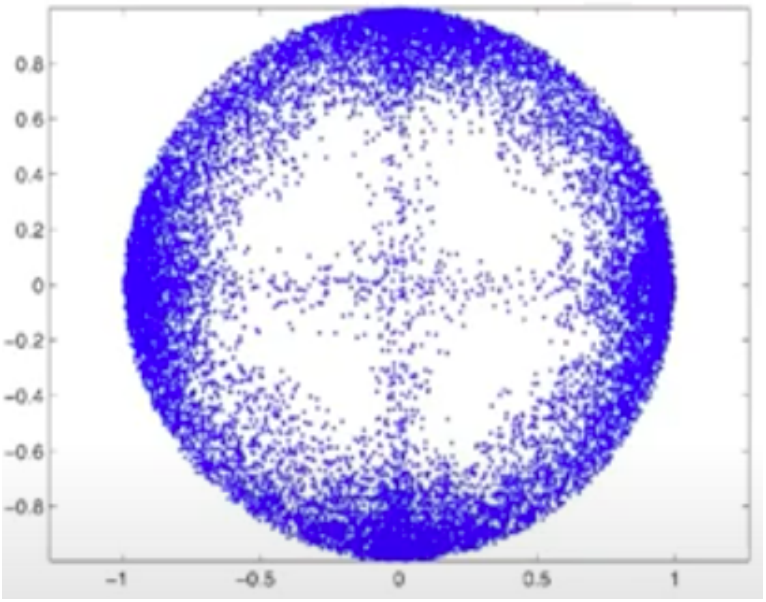

Figure 10.6: *Using PCA on the high and medium density part of the data.*

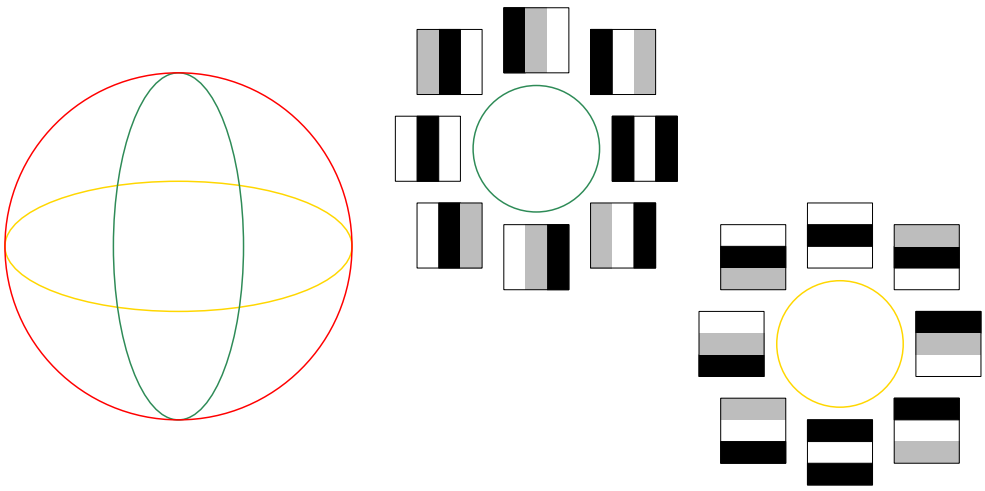

Figure 10.7: *The interpretation of the high and medium density part of the data.*

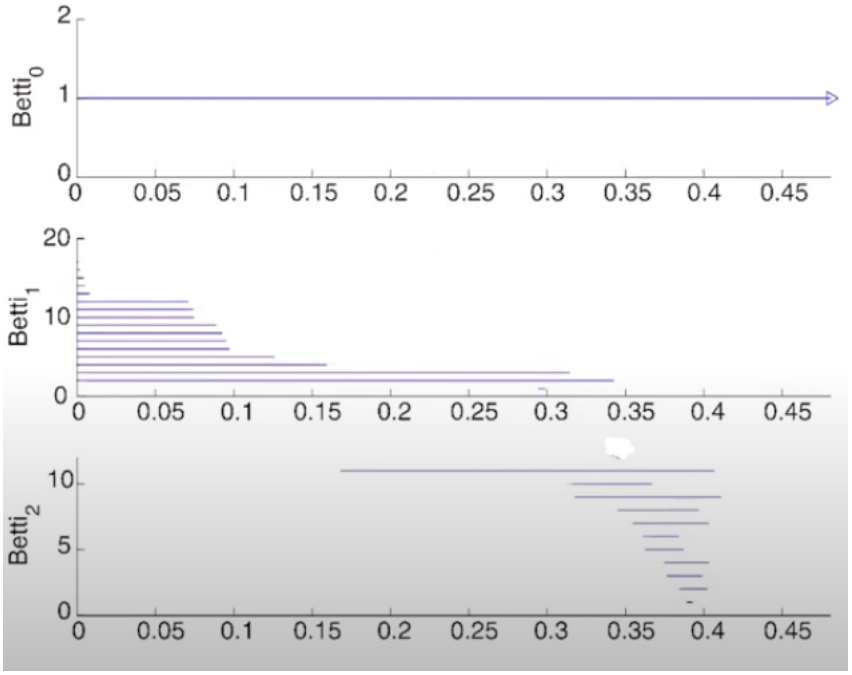

Figure 10.8: *The barcodes for the entire data.*

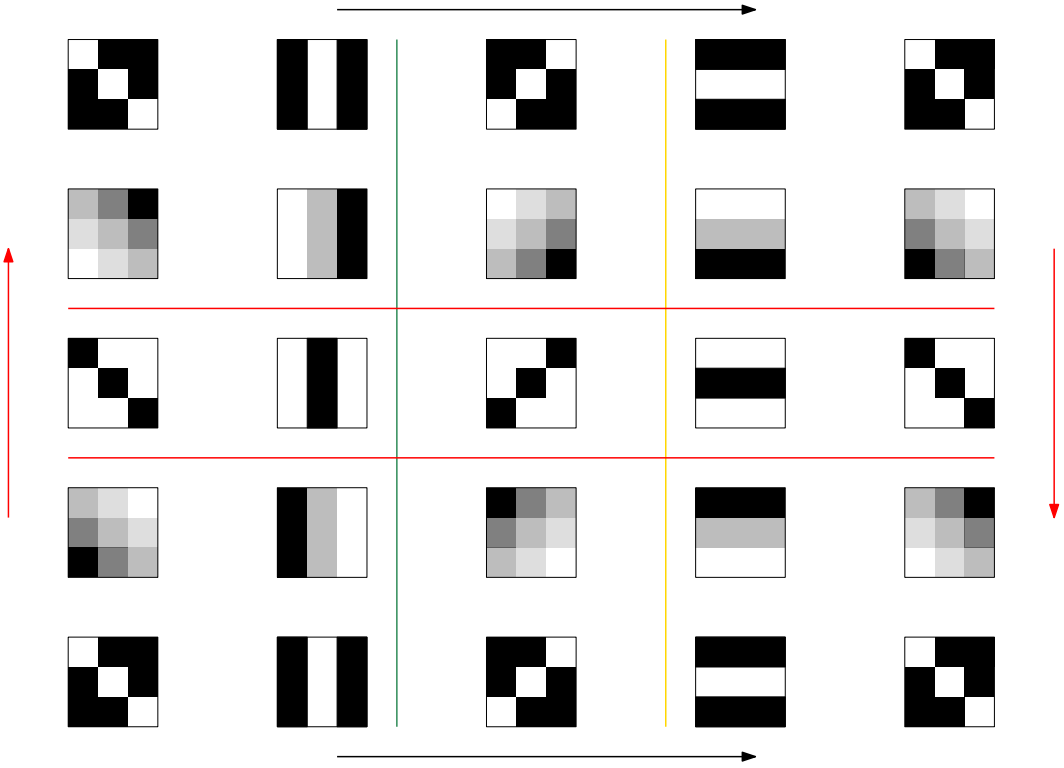

Figure 10.9: *The interpretation of the entire data.*

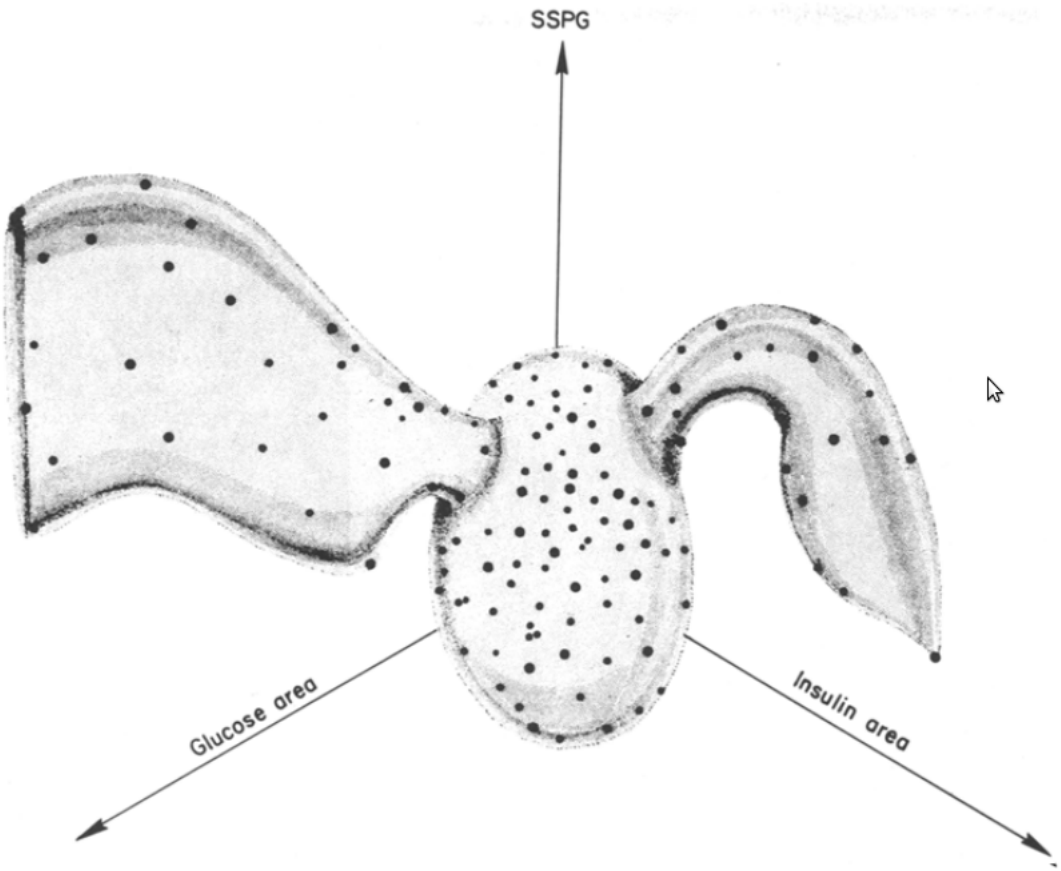

Figure 10.10: *The classical illustration of the diabetes data set.*

measured per participant: age, relative weight, fasting plasma glucose, area under the plasma glucose curve for the three hour glucose tolerance test (OGTT), area under the plasma insulin curve for the (OGTT), and steady state plasma glucose response. After applying various classical methods, the original study came up with a picture containing a blob of healthy people, with two strains coming out of this blob, called type 1 and type 2 diabetes [4]. See Figure 10.10 for their illustration. The same behaviour can be automatically detected using Mapper, as showcased in [6]. As the filter function they used a density estimator for each point. They created two different outputs, one for 3 and one for 4 intervals, always with a 50% overlap. In the output of Mapper depicted in Figure 10.11 you also see areas of large density, with 2 flares going out of it.

The Mapper approach has later also been applied to genomic data from a breast cancer study [7] by Nicolau, Levine and Carlsson [5]. This data is very high dimensional, each patient defines a data point in $\mathbb{R}^{262}$. As a filter function they used the distance to some baseline healthy tissue. The output presented in their paper is depicted in Figure 10.12. The strain labelled *c-MYB*$^+$ *tumors* corresponds to a type of tumor which has previously not been classified, where the other strains confirm the strains of breast cancer known previously.

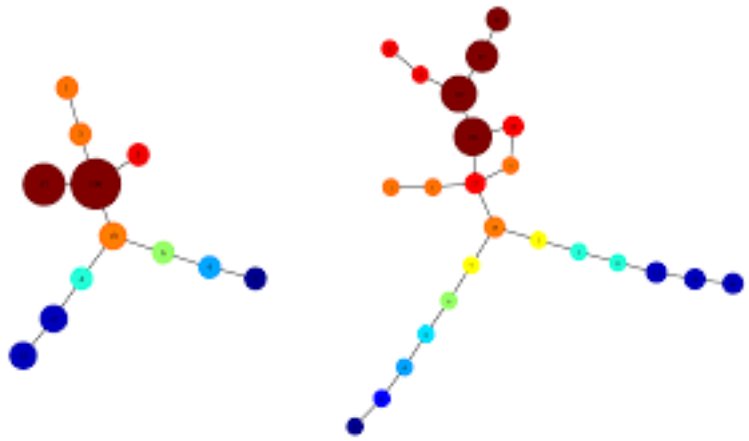

Figure 10.11: *The output of Mapper on the diabetes data set.*

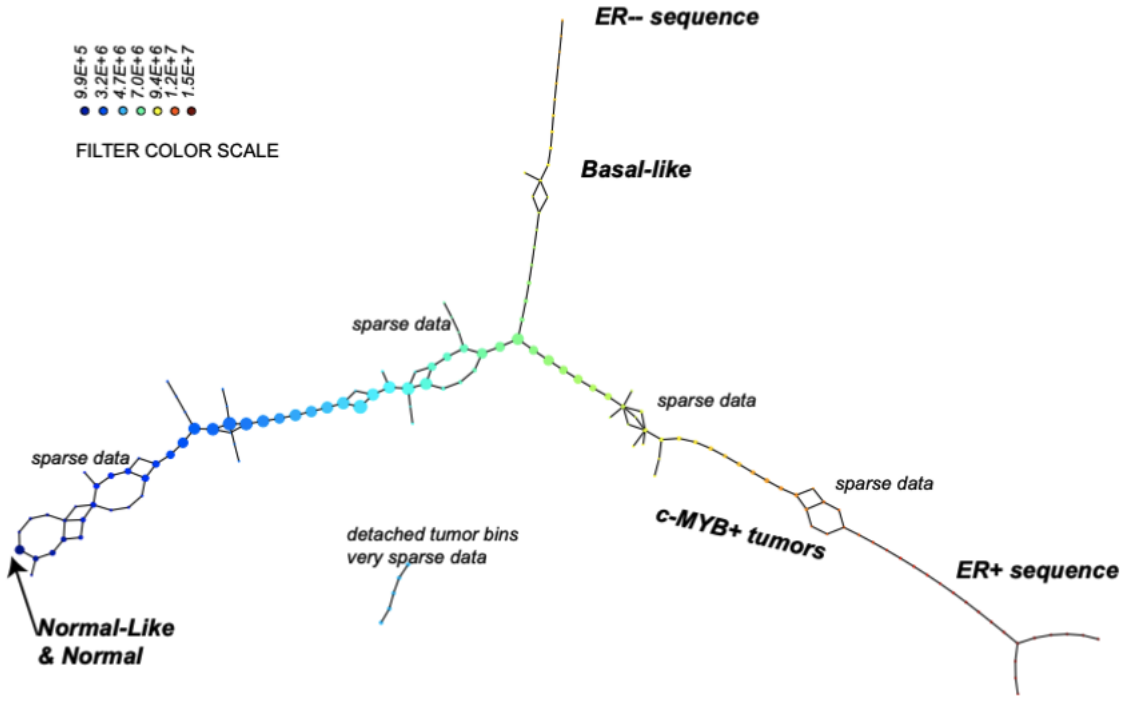

Figure 10.12: *The output of Mapper on the breast cancer data set.*

Topological data analysis has been used in many studies in medicine. TDA seems to excel in these high-dimensional datasets since many features seem to be the result of higher-order interactions of different coordinates. TDA can also deal with much smaller data sets than other approaches such as deep neural networks, that need tons of data points to train, validate, etc. In medicine, studies often have very few data points due to the large monetary cost, workload, and also ethical questions involved with data gathering.

10.3 Time Series

Time series is data given as a sequence of points x_t in some metric space X, where t is a (discrete) variable. The goal in analyzing time series is often analyzing and finding periodic behavior in the time series.

We can embed time series in some higher-dimensional space, by always considering a sequence of $l+1$ consecutive points, i.e., $\mathcal{I}_l := \{(x_{t_i}, x_{t_{i+1}}, \ldots, x_{t_{i+l}})\} \subset X^{l+1}$. The idea is that periodicity in the time series translates to loops (1-dimensional holes) in $\mathcal{I}_l$. As an example, consider $x_t = \sin(\frac{\pi}{4}t)$, and consider $l = 1$. Then, we get a loop in $\mathbb{R}^2$. If we however consider $x_t = t$, $\mathcal{I}_1$ is just a straight line.

This approach has been used by Vejdemo-Johansson, Pokorny, Skraba and Kragic [9] for analyzing motion capture data, where cameras track the location in 3-space of certain points on a human body marked by physical markers. The data considered are six seconds long recordings of movement such as boxing, see Figure 10.13 for some stills. Every data point in this set is roughly 70-dimensional. Using PCA to project onto 2 dimensions, some loops can be seen but we cannot really distinguish different loops, see Figure 10.14 (left). Computing persistent homology it can be seen that there are six different loops that persist a long time, see Figure 10.14 (right). Looking back at the input data, these six different loops can be mapped to six different boxing movements in the recording.

10.4 Politics

The data considered in this application, done in [2], is gathered from the sessions of the US house of representatives through the year 2010. For every vote, we set $x_i = 1$ if the member i says Yes to the vote, -1 for No, and 0 otherwise. Doing this for every of the 447 representatives in all 664 votes this gives 664 data points in $\mathbb{R}^{447}$. Using PCA to project to two dimensions and coloring the data by splitting the members into the two parties we get the picture depicted in Figure 10.15. We can see that there are four main "corners" in which issues lie: the ones that get bipartisan support, those that get bipartisan rejection, those that are supported by republicans and rejected by democrats, and those where it is the opposite way. We can also see that the corner with bipartisan rejection is very sparsely populated, since such issues rarely make it far enough to be voted on. Further there are few votes that lie in between two of these corners, and almost none that lie

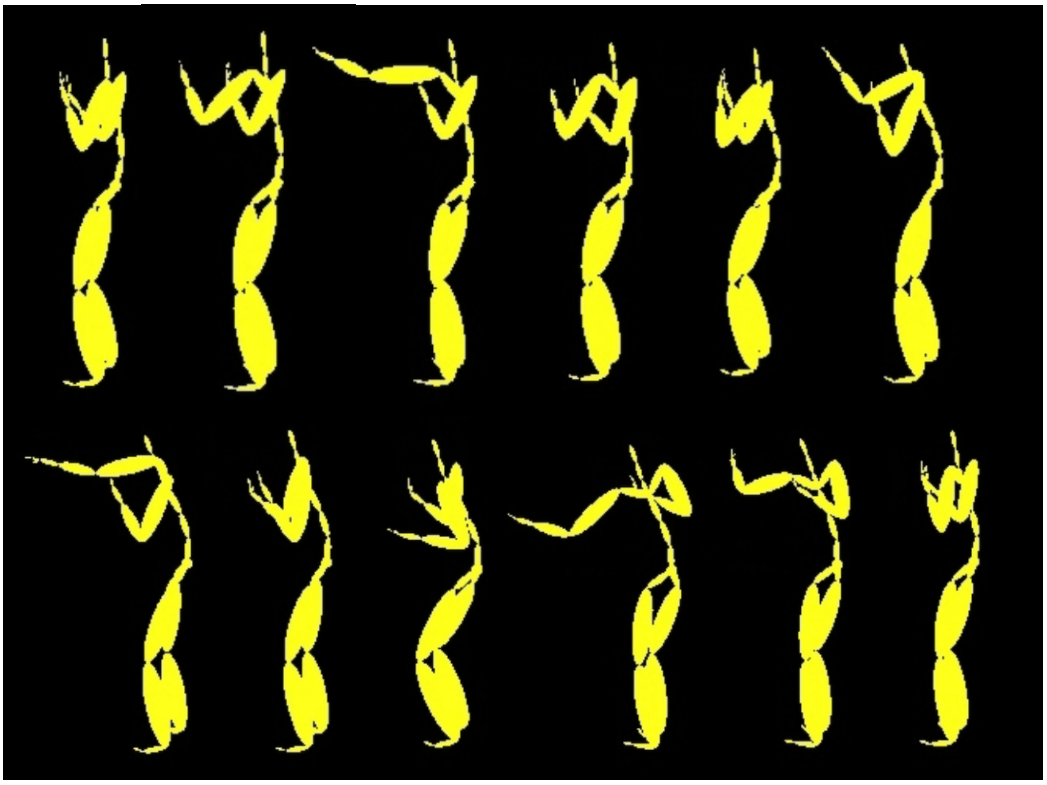

Figure 10.13: *Some stills from motion capture data of a boxer.*

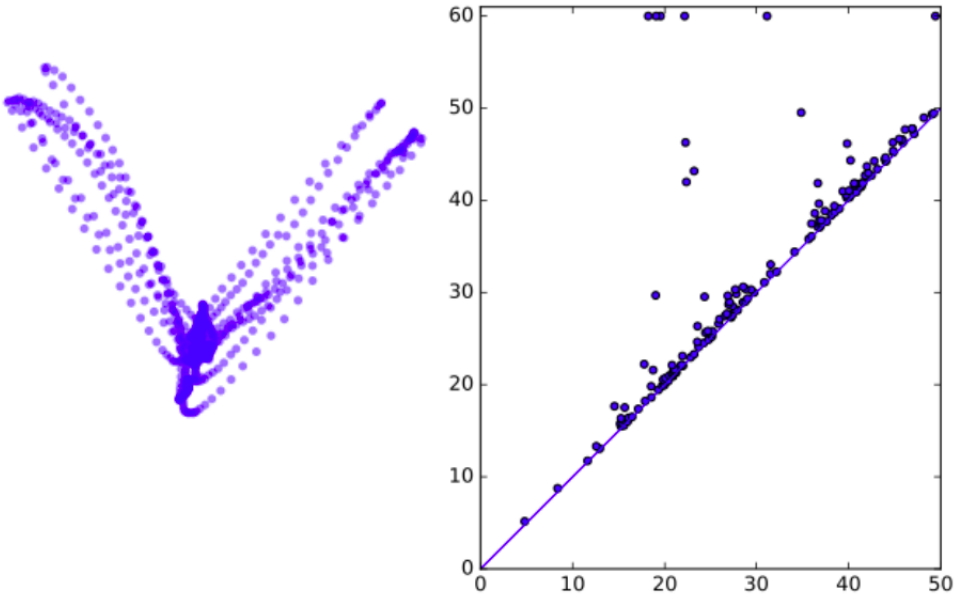

Figure 10.14: *Left: PCA of the data set obtained from the boxing recording. Right: the persistent homology of the same data.*

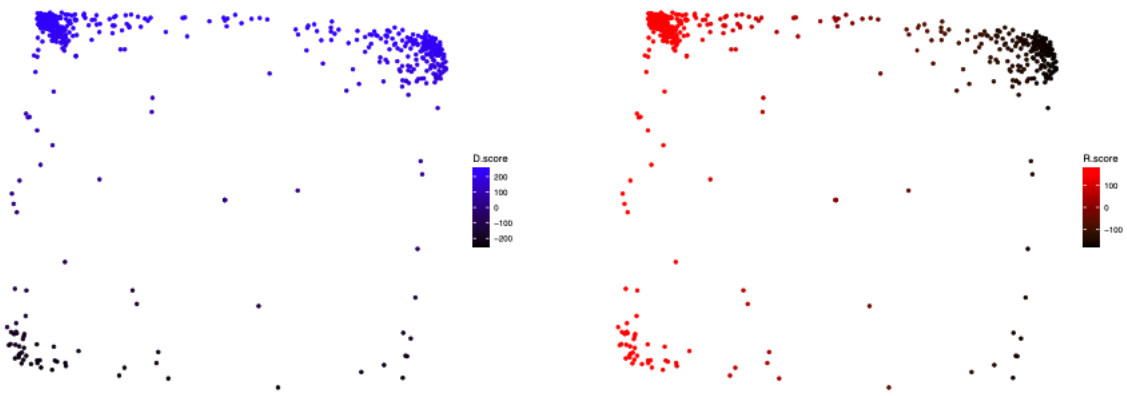

Figure 10.15: *PCA of the voting data from the US house of representatives, colored by parties.*

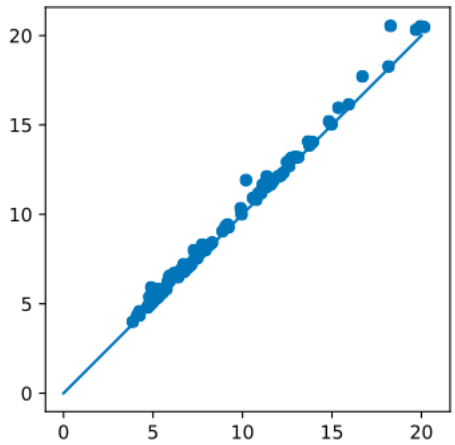

Figure 10.16: *Persistence diagram in dimension 1, computed on the entire voting data.*

in the middle of all four corners. There are however some: in particular, three of the recorded votes where Quorum calls, where both answers (Present/Not Present) are coded as 0, so there are at least 3 points exactly in the middle of the square.

Computing the 1-dimensional persistence diagram, we would hope to recognize the phenomenon of excluded middle by having a 1-dimensional homology class with long persistence, corresponding to the boundary of the square. However, as we can see in the diagram depicted in Figure 10.16, this is not the case. What is going on? If you think about the process of growing balls, it becomes apparent that already a single point in the center of the square is enough to "fill the hole" much faster than we would like. As you can see in Figure 10.15, there are several points in the middle of the square, and it is these points that make the persistence of any 1-dimensional homology class short. This is a general issue of persistent homology: it is very fragile to outliers in the data.

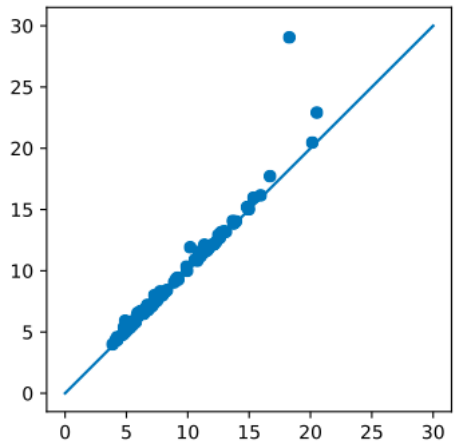

Figure 10.17: *Persistence diagram in dimension 1, computed on 99% of the voting data.*

One general approach to overcome this issue in many application is the following: compute the *density* for each data point (e.g., using Gaussian kernel density estimation or any other standard method) and only keep the points with highest density, e.g., remove the 1% with lowest density. Then, compute persistent homology only on the remaining, dense points. Doing this on the voting data, we get the diagram depicted in Figure 10.17. Here we can clearly see a single homology class with long persistence, just as we would expect.

10.5 Shape Segmentation

Consider a surface in $\mathbb{R}^3$ given as a triangular mesh, that is, by points, edges, and triangles. The goal of *shape segmentation* is to label different parts of the surface according to what part they are. For example, given a model of a human body, we want to segment it into categories such as "head", "torso" or "upper arm". While this is a well-studied problem in computer graphics, here we sketch a topological approach, described in [3].

We can pick some point on the body, and start growing a ball around it, using the geodesic distance (length of the shortest path along the surface). On the resulting filtration, we can perform persistent homology in dimension 1. If we do this for a point on the palm of a hand, for example, we get a 1-dimensional hole for every finger, see Figure 10.18. If we do this for a point on a finger, the persistence diagram looks very different, see Figure 10.19. We can then classify the persistence diagrams to segment the shape. But, how can we do this? We need to be able to insert persistence diagrams into classical ML pipelines. For this there are several approaches, and we discuss some of them in the next section.

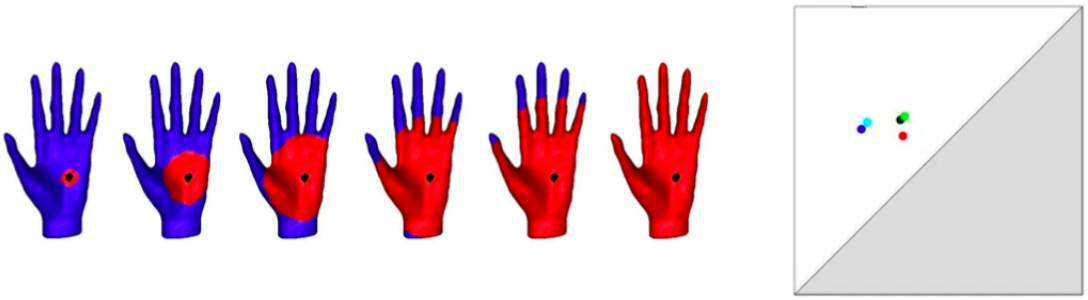

Figure 10.18: *Persistence diagram for a point on the palm.*

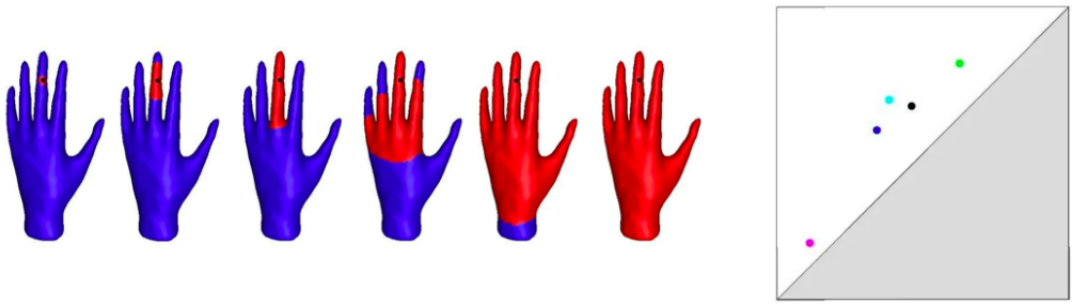

Figure 10.19: *Persistence diagram for a point on a finger.*

10.6 TDA in ML

Many machine learning pipelines require input points to be in Euclidean space (and not just any metric space, which the persistence diagrams would already fulfill), or in the case of kernel methods, at least in some space that has an inner product, also known as a *Hilbert space*.

There are many ways to turn persistence diagrams into elements of such metric spaces. These methods are also called *vectorizations*. In this section we introduce three such methods. On https://persistent-homology.streamlit.app you can see more examples and play around with various vectorization methods.

Persistence Statistics For the persistence statistics, as the name suggests, we just summarize some statistical values of the persistence diagrams. In particular, for the birth times b, the death times d, the interval midpoints $\frac{b+d}{2}$ and the interval lengths $d - b$ we record

- the mean,
- the standard deviation,
- the median,
- the interquartile range,

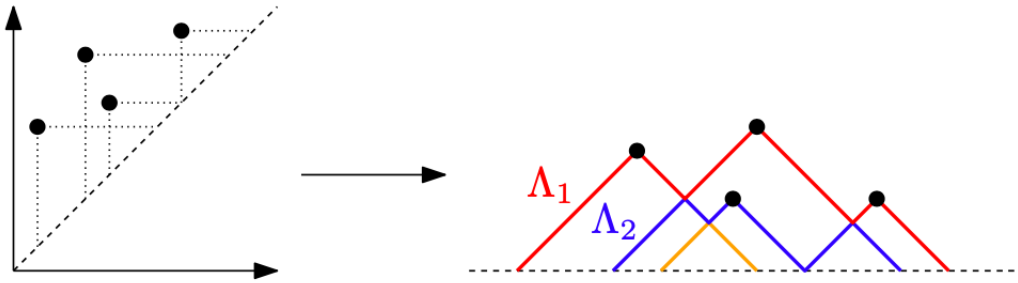

Figure 10.20: *The persistence landscape of a persistence diagram.*

- the full range and
- four percentiles (10%, 25%, 75%, 90%).

Finally we also record the number of bars and the entropy. All in all, we record 38 numerical values, and we thus represent a persistence diagram as a point in $\mathbb{R}^{38}$.

Persistence Landscapes The main intuition behind persistence landscapes is the following: for each point in the persistence diagram draw a horizontal and a vertical line segment to the diagonal. Flipping this arrangement of line segments such that the diagonal is horizontal, we get something that looks like a mountain range, hence the name landscape. We can also interpret this as a set of piece-wise linear functions by looking at *envelopes*: a point on one of the segments is on the k'th envelope of the arrangement if there are $k-1$ segments strictly above it. Each such envelope is now a piece-wise linear function. See Figure 10.20 for an illustration.

More formally, let $D = \{(b_1, d_1), \ldots, (b_n, d_n)\}$ be a persistence diagram with finitely many off-diagonal points. Each off-diagonal point (b_i, d_i) gives rise to a triangle whose boundary is defined by the points

$$\{(t, \min[t - b_i, d_i - t]_+) \mid t \in \mathbb{R}\},$$

where $\min[a, b]_+ := \max(0, \min(a, b))$.

The persistence landscape of D is now a function $\lambda_D : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ defined by

$$\lambda_D(k, t) := \text{ k'th largest value of } \min[t - b_i, d_i - t]_+ \text{ for } i \in \{1, \ldots, n\}.$$

For each k, we have that $\lambda_D(k, \cdot)$ is a piece-wise linear function. Recall that on functions, we have the following p-*norms*:

$$\|f\|_p := \left(\int_{\mathbb{R}} |f(x)|^p \, dx \right)^{1/p}.$$

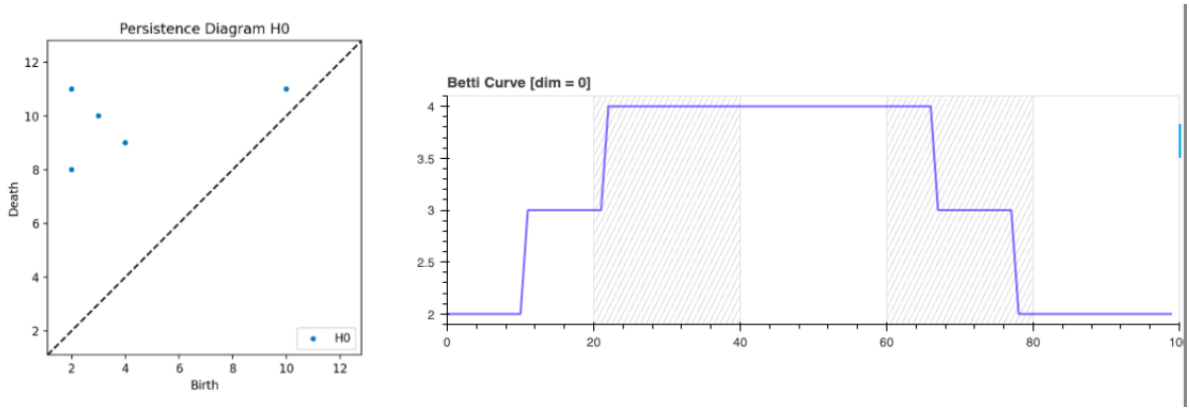

Figure 10.21: *The Betti curve of a persistence diagram.*

We can extend this to a norm on our set of piece-wise linear functions:

$$\|\lambda_D\|_p := \left(\sum_{k=1}^{\infty} \|\lambda_D(k, \cdot)\|_p \right)^{1/p}.$$

In particular, for $p = 2$ this is a Hilbert space and thus usable for many machine learning pipelines. We can further define the following distance measure on persistence diagrams, called the p-*landscape distance* Λ_p:

$$\Lambda_p(D_1, D_2) := \|\lambda_{D_1} - \lambda_{D_2}\|_p.$$

Taking $p = \infty$ we get the following

Theorem 10.1. *Let D_1 and D_2 be persistence diagrams with finitely many off-diagonal points. Then $\Lambda_\infty(D_1, D_2) \leqslant d_b(D_1, D_2)$.*

Betti Curves An easier way to get a piece-wise linear (even piece-wise constant) function from a persistence diagram is through Betti curves: the Betti curve of a persistence diagram is the function $\beta : \mathbb{R} \to \mathbb{N}_{\geqslant 0}$ which assigns each time t the current Betti number of the filtration. This is a piece-wise constant function and thus the space of all Betti curves is a Hilbert space with the standard 2-norm. A Betti curve still captures all the births and deaths, but throws away the pairing between them, see Figure 10.21 for an illustration.

Questions

39. *How can we use TDA to analyze the space of image patches?* Discuss the structure of the relevant data set and how persistent homology can be applied.

40. *Why does persistent homology sometimes fail to capture voids in the data?* Illustrate the problem with the voting data from the US house of representatives and explain how it can be solved in practice.

41. *How can persistence diagrams be used in machine learning pipelines?* Explain the following vectorizations: persistence statistics, persistence landscapes and Betti curves.

References

[1] Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian, On the local behavior of spaces of natural images. *International journal of computer vision*, 76, (2008), 1–12.

[2] Gunnar Carlsson and Mikael Vejdemo-Johansson, *Topological Data Analysis with Applications*, Cambridge University Press, 2021.

[3] Mathieu Carrière, Steve Y. Oudot, and Maks Ovsjanikov, Stable Topological Signatures for Points on 3D Shapes. *Computer Graphics Forum*, doi:10.1111/cgf.12692.

[4] Rupert G Miller, Discussion: Projection Pursuit. *The Annals of Statistics*, 13/2, (1985), 510–513.

[5] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108/17, (2011), 7265–7270.

[6] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.

[7] Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton et al., A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347/25, (2002), 1999–2009.

[8] J Hans Van Hateren and Arjen van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265/1394, (1998), 359–366.

[9] Mikael Vejdemo-Johansson, Florian T Pokorny, Primoz Skraba, and Danica Kragic, Cohomological learning of periodic motion. *Applicable algebra in engineering, communication and computing*, 26/1, (2015), 5–26.