Projects in Topological Data Analysis 2024

Contents

1 General information

The main idea of the course is that students from different universities collaborate on projects in topological data analysis. These projects can be of theoretical nature, e.g. the development or analysis of some algorithm or applied, e.g. using topological data analysis on concrete data sets.

The course will start with two lectures introducing some topics that are relevant to the proposed projects. Then, there is an official kick-off meeting, in which the problems will be introduced. The assignment of the groups will be made in the days after the kick-off meeting, so that the groups can start working on their projects as soon as possible. For the main part of the course, the groups work on their project individually, guided by the two mentors they are assigned. Once during the course, each group will also present their progress to the other mentors, collecting some more feedback from different people. At the end of the course, each group hands in a written report where they summarize their work and present their findings.

1.1 Important dates

All events will start at 17:00 Zürich/Eindhoven/Fribourg time; 11:00 Poughkeepsie/Notre Dame time; 8:00am Eugene time.

September 3, 2024	Intro lecture by Patrick Schnider on persistent homology and Mapper
September 10, 2024	Intro lecture by Tao Hou and Dev Sinha on cohomology and Hodge decomposition
September 17, 2024	Kick-off meeting
During the course	Groups work on their projects individually, guided by the mentors
December 17 and 18, 2024	Deadline for handing in the final reports and meeting for final presentations

1.2 Formalities for the final reports and presentations

The final reports should be structured like a research paper in the area. In particular, they should contain an overview of the relevant literature, clearly highlight the novel contributions and precisely describe the technical content. There is no formal page limit, each group can use as many or few pages as they need to write their paper in a way that they find appropriate for presenting their work. For questions about the structure and contents of the reports, the mentors are a valuable resource of help.

In the final meeting, each group presents their paper in a 20 minute talk, followed by 5 minutes of questions from the audience. This talk can be given by a single group member or by several people. The goal of this talk is, that the other groups get to see your results, so it should be prepared with the peers as a target audience.

After the final meeting, all mentors will decide on the final grades for each individual person taking the course. For this, the work during the semester, the final report as well as the presentation will be taken into account. Finally, the grade will be converted to the grading system of the local university by the local mentor.

1.3 Additional information and communication channels

Additional information can be found on the course webpage. In particular, we intend to maintain a list of sources on all aspects topological data analysis that are relevant for the projects.

Each group will have access to a zoom room for their meetings, and an overleaf file for their write-ups. For sharing of code and other documents, each group can request access to an individual gitlab page (gitlab is a github-like clone provided by ETH Zürich). Finally, the communication that is important for all groups and mentors will take place on a discord server.

2 List of Projects

2.1 Persistent Homology of Gerrymandering

The area of redistricting in the United States has gotten much attention of late, as many groups work to quantify and evaluate plans as well as designing tools and techniques which can evaluate the concept of "fairness" in this domain. Motivated by recent work that uses tools from topological data analysis in the domain of computational redistricting [1], one open area is to apply tools more widely in this domain. In particular, this paper only computes bottleneck distances given fixed voting data, but does not consider other metrics such as optimal transport.

Problem 1. *How does the data from [1] behave under different metrics?*

It might also be of interest to study this data using Reeb graphs or other shape descriptors from topological data analysis, rather than simple persistence diagrams.

Problem 2. *Can we find other types of structures in the data from [1]?*

In a different direction, using the framework in [1], one could attempt to locate areas in a state that are split differently in party-biased ensembles or whose splitting correlates with party advantage.

Mentors: Erin Chambers, Anna Schenfisch, Tao Hou

References

[1] Moon Duchin, Tom Needham, and Thomas Weighill. The (homological) persistence of gerrymandering, 2020.

2.2 The Shape of the Swiss Railway System

Both a country as well as the traffic systems in a country are inherently geometric. However, the traffic system often has different geometric features than the underlying country. For example, traffic systems are usually much denser in areas where many people live. In contrast, there are often stretches of land that are only sparsely inhabited, taking up a lot of space of the country, but not providing much to the geometry of the traffic system.

In this project, the goal is to analyze geometric and topological features of the Swiss Railway system. The relevant data for this is publicly available, and it can be either treated as a finite metric space or as a network. Thus, using methods from topological data analysis for finite metric spaces or for graphs we can analyse the *intrinsic* and *extrinsic* shape of Switzerland from the perspective of travellers.

Mentors: Bastian Rieck, Tim Ophelders

2.3 Inverse Problem with Mapper Graphs

The Mapper algorithm is a popular tool for visualization and data exploration in topological data analysis. The recent paper "Any graph is a Mapper graph" [1] investigates an inverse problem for the Mapper algorithm: Given a dataset X and a graph G, does there exist a set of Mapper parameters such that the output Mapper graph of X is isomorphic to G? Two constructions are provided that affirmatively answer this question. Some natural follow-up questions are:

- What are some other constructions?
- What are some constructions when we add constraints (e.g. reference space is $\mathbb{R}$, or when we use a particular type of clustering algorithm)?
- Is there anything we can say about "how big" the set of Mapper parameters is that work for a particular graph and dataset.
- For the construction that maps into convex sets in $\mathbb{R}^3$, what is the map that allows for the largest extension?
- Write code that constructs convex sets in $\mathbb{R}^3$ whose nerve is G; starting with subsets of $\mathbb{R}^4$ is easier.

Mentors: Robin Belton, Enrique Alvarado

References

[1] Enrique G Alvarado, Robin Belton, Kang-Ju Lee, Sourabh Palande, Sarah Percival, Emilie Purvine, and Sarah Tymochko. Any graph is a mapper graph, 2024.

Figure 1: An example of a political profile.

2.4 Holes in Swiss Politics?

For many elections in Switzerland, the webpage *smartvote* [1] publishes a questionnaire that both voters and candidates can fill out. Based on the questions, a political profile is created that includes scores in 8 axis. This profile is visualised in a spider diagram, see Figure 1. Based on this profile, smartvote gives voters a ranked list of candidates that most closely agree with their profile, facilitating the choice between the many candidates and parties in Swiss elections.

Smartvote has provided us with the data of all candidates of last year's Swiss Federal Elections. From a perspective of data analytics, each political profile can be interpreted as a data point in $\mathbb{R}^8$. It is to be expected that some of the 8 axis are correlated and that these data points form some manifold. The goal of this project is to try and understand this manifold.

Problem 3. *What are some topological properties of this "space of candidates"? Does it have any non-trivial homology? What is its dimension? Is there any curvature? Does it have singularities?*

Mentors: Patrick Schnider, Simon Weber

References

[1] smartvote. `https://www.smartvote.ch`, 2024.

2.5 Insightful Applications of Hodge Decomposition / Harmonic Representatives

Consider the p-th cohomology group $H^p(K;\mathbb{R})$ of a simplicial complex K over the coefficient $\mathbb{R}$, where each p-cochain is a linear map $f : C_p(K) \to \mathbb{R}$. The *Hodge p-Laplacian* $\Delta_p : C^p(K;\mathbb{R}) \to C^p(K;\mathbb{R})$ is a linear map defined as:

$$\Delta_p : \delta_{p-1}\delta_{p-1}^* + \delta_p^*\delta_p,$$

where δ_p, δ_{p-1} are the coboundary maps and $\delta_p^*, \delta_{p-1}^*$ are their adjoints.

The *Hodge decomposition* is then the following:

$$C^p(K;\mathbb{R}) = \mathrm{im}(\delta_p^*) \oplus \ker(\Delta_p) \oplus \mathrm{im}(\delta_{p-1}).$$

We also have that $\ker(\Delta_p)$ is isomorphic to $H^p(K;\mathbb{R})$ (and hence $H_p(K;\mathbb{R})$) and that each cocycle in $\ker(\Delta_p)$ is called *harmonic* which minimizes the norm in the corresponding coset in $H^p(K;\mathbb{R})$.

The special cases of the things defined above are probably more well-known: when K is a graph, Δ_0 is the *graph Laplacian*, and when K is two-dimensional, Δ_1 is the *graph Helmholtzian*. Moreover, it is known that the cochains are discrete analogues of differential forms on manifolds and the coboundary operators are discrete analogues of exterior derivatives. Hence, Δ_1 is also

$$\Delta_1 : -\mathrm{grad}\,\mathrm{div} + \mathrm{curl}^*\mathrm{curl}.$$

See [4] for a more detailed introduction to Hodge decomposition and harmonic representatives.

In several works (e.g., [1,3,4,6]) the authors describe some applications of Hodge decomposition. One of them is as follows: Consider a set V of objects to be ranked (such as movies). Here, we are given a *pairwise ranking*

$$X : E \to \mathbb{R}$$

which can be considered a 1-cochain, where $X(u, v)$ denotes how much the object u is favored over v. Applying the Hodge decomposition in dim 1 gives:

$$X = \mathrm{grad}(f) + X_H + \mathrm{curl}^*(\Phi),$$

where $f : V \to \mathbb{R}$ is a global scoring on the objects and X_H is harmonic. Roughly speaking, $\mathrm{grad}(f)$ measures the consistency of the pairwise ranking and X_H, $\mathrm{curl}^*(\Phi)$ measure the (global and local) inconsistency of it.

Problem 4. *Can we find other 'natural' and insightful applications of Hodge decomposition as above?*

On a separate note, [2] (and related work [5]) describe applications of harmonic representatives (e.g., see below:)

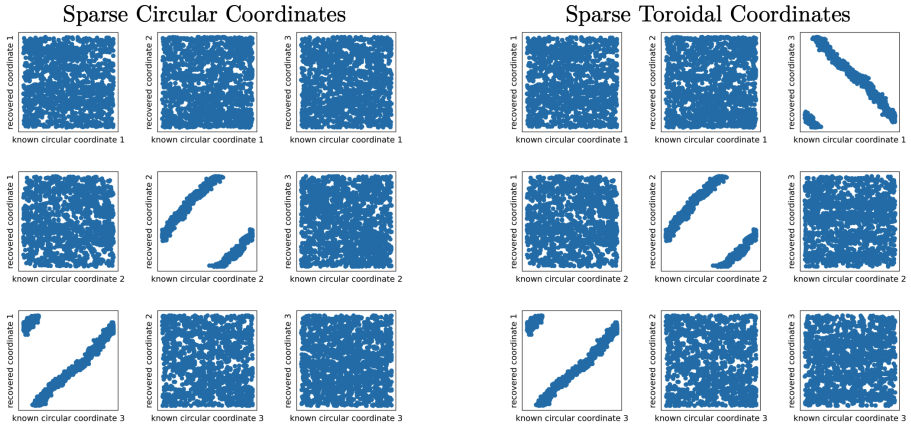

Figure 12 Recovered versus known circular coordinates using the Sparse Circular Coordinates Algorithm and the Toroidal Coordinates Algorithm.

Figure 2: Taken from [5]

Problem 5. *Can we find other interesting applications of harmonic cycles?*

A more open one (note: there are already some works in the TDA community for this which are not listed here):

Problem 6. *How can Hodge decomposition / harmonic representatives be integrated into TDA in other ways?*

Mentors: Tao Hou, Anna Schenfisch, Dev Sinha

References

[1] Michael J. Catanzaro and Brantley Vose. Harmonic representatives in homology over arbitrary fields. *Journal of Applied and Computational Topology*, 7(3):643–670, 2023.

[2] Vin De Silva and Mikael Vejdemo-Johansson. Persistent cohomology and circular coordinates. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 227–236, 2009.

[3] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

[4] Lek-Heng Lim. Hodge laplacians on graphs. *Siam Review*, 62(3):685–715, 2020.

[5] Luis Scoccola, Hitesh Gakhar, Johnathan Bush, Nikolas Schonsheck, Tatum Rask, Ling Zhou, and Jose A Perea. Toroidal coordinates: Decorrelating circular coordinates with lattice reduction. *arXiv preprint arXiv:2212.07201*, 2022.

[6] Ronald Koh Joon Wei, Junjie Wee, Valerie Evangelin Laurent, and Kelin Xia. Hodge theory-based biomolecular data analysis. *Scientific Reports*, 12(1):9699, 2022.

2.6 The Space of Soccer Passes

In the sport of soccer[1], one of the most common events is that of a *pass*, where one player kicks the ball to another player. A pass is determined by several factors, e.g. the position of the pass player on the field, the direction and velocity of the pass, whether the pass was on the ground or in the air, whether it was successful, etc. Each pass can thus be seen as a point in some high-dimensional space X.

There are several data sets that collect this data for all passes played during a match or even during a championship. Two publicly available data sets are the StatsBomb data set [2] or the data from the Soccer Data Challenge initiative [1]. More data and some projects can be found on Edd Websters github page [3].

The goal of this project is to use topological data analysis to study soccer passes. It is to be expected (and is also indicated by projection to $\mathbb{R}^2$) that the passes that actually occur during games lie on some low-dimensional subspace P of X.

Problem 7. *What can we say about P? Is it connected? What is its dimension? What is its homology? Is it a manifold?*

Depending on tactics, different teams likely play different passes.

Problem 8. *Can the space of passes be used to classify different teams in a tournament? Does the space of passes of a single team change significantly between different games?*

Mentors: Patrick Schnider, Tim Ophelders

References

[1] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):236, 2019.

[2] StatsBomb. Statsbomb open data. `https://github.com/statsbomb/open-data/blob/master/README.md`, 2022.

[3] Edd Webster. Edd webster football analytics. `https://github.com/eddwebster/football_analytics`, 2023.

[1] Also known as football.

2.7 Dimensionality Reduction for Manifolds

Dimensionality reduction is an important technique in data science, and thus many different methods have been proposed, ranging from classical projection tools like PCA to deep learning techniques combined with topological ideas, e.g., topological autoencoders [1]. Of course, all of these methods will lose some information, but generally different methods retain different types of information. Depending on the application, it is thus important to choose the correct method of dimensionality reduction.

A common assumption for many machine learning tasks is the so-called *manifold hypothesis*, which assumes that the underlying data is sampled from a manifold. Thus, it is a natural question to understand the behavior of different dimensionality reduction methods for data sampled from manifolds.

In this project, the goal is to experimentally compare different dimensionality reduction methods for data sampled from different manifolds, in particular high-dimensional manifolds that appear in many contexts, such as $SO(n)$.

Mentors: Bastian Rieck, Robin Belton

References

[1] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7045–7054. PMLR, 13–18 Jul 2020.

2.8 Persistence of generalized density functions

One potential way to infer the shape of a data set in $\mathbb{R}^n$ is to define some appropriate density function on $\mathbb{R}^n$ which intuitively should be higher in areas where the data points are dense, and then computing the persistent homology of the corresponding superlevelset or sublevelset filtration. Ideally, the density function would further have some nice properties, e.g., that the computation of persistent homology can be done efficiently, or that we get some stability, e.g. that the bottleneck distance of two persistence diagrams is bounded by the Hausdorff distance of the two data sets.

One example of such a function is the one which, given a set P of data points in $\mathbb{R}^n$ assigns to a point $x \in \mathbb{R}^n$ the distance to its closest point in P, i.e., $f(x) := \min_{p \in P} d(x, p)$. You can convince yourself that the persistence diagram $Dgm_{f,P}$ of the corresponding sublevelset filtration is exactly the persistence diagram of the Čech filtration of P. In particular, from the stability theorem for Čech filtrations we get that for two different data sets P and Q, the bottleneck distance of the persistence diagrams is bounded from above ba the Hausdorff distance of P and Q, that is,

$$d_b(Dgm_{f,P}, Dgm_{f,Q}) \leqslant d_H(P, Q).$$

The goal of this project is to investigate for which functions we get such a stability theorem. More formally, a *generalized density function* is a function that takes as input a data set P in $\mathbb{R}^n$ and an additional point $x \in \mathbb{R}^n$ and assigns a real value. In particular, for every P we get a function $\mathbb{R}^n \to \mathbb{R}^n$.

Problem 9. *For which generalized density function do we have a stability theorem?*

There are of course some trivial ones, e.g. constant functions or any function not depending on P, but as we have seen above there are also more interesting examples.

Mentors: Patrick Schnider, Enrique Alvarado