Projects in Topological Data Analysis

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 three 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 16:15 Zürich/Einhoven time and 9:15am Chicago/St. Louis time.

August 29, 2023	Lecture on directional transforms (Erin Chambers)
September 5, 2023	Lecture on software packages (Tao Hou)
September 12, 2023	Lecture on sequences in topology (Patrick Schnider)
September 19, 2023	Kick-off meeting
During the course	Groups work on their projects individually, guided by the mentors
December 12, 2023	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 seminar. 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 will get 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, Patrick Schnider

References

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

2.2 Mayer-Vietoris for bi-chromatic data sets

Assume you are given a data set where each data point is colored either red or blue. For the Čech complex, this gives a 2-coloring of its vertices. Further, this two coloring induces a natural cover of the complex with two subcomplexes: one is the complex R of all faces contained in a maximal face that contains a red vertex, and the complex B is the same for the blue vertices. The intersection of R and B is the complex induced by maximal faces with vertices of both colors.

Recall the Mayer-Vietoris sequence

$$\dots H_{k+1}(X \cup Y) \to H_k(X \cap Y) \to H_k(X) \oplus H_k(Y) \to H_k(Y) \to H_k(X \cup Y) \to \dots$$

The pair R, B induces a Mayer-Vietoris sequence, and by considering a filtration of the underlying complex, we get a sequence of Mayer-Vietoris sequences.

Mayer-Vietoris sequences in persistent homology have already been studied in more general settings [1, 2]. The goal of this project is to find algorithms to compute the relevant sequences in the setting described above.

Problem 3. *Can the sequence of Mayer-Vietoris sequences or the ranks of the involved groups be computed efficiently?*

Some variants that could be interesting and more efficient from a computational point of view are, if the underlying data points have some geometric constraints, e.g. that the blue points can be separated by a hyperplane from the red points.

Mentors: Tao Hou, Patrick Schnider

References

[1] Barbara Di Fabio and Claudia Landi. A mayer–vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. *Foundations of Computational Mathematics*, 11:499–527, 2011.

[2] Álvaro Torras Casas. Distributing persistent homology via spectral sequences, 2020.

2.3 The structure of the cosmic web

The cosmic web is the structure emerging from the positions of galaxies in space. The Bolshoi data set is a particle simulation of the cosmic web [1]. Looking at local densities, this structure seems to form a cell-complex, consisting of 0-dimensional *clusters*, 1-dimensional *filaments*, 2-dimensional *walls*, and 3-dimensional *voids*. This structure has already been studied using persistent homology [2]. In reality, however, voids are not completely empty, so traditional persistent homology is thrown off by outliers.

Problem 4. *Can we adapt persistent homology to work around this?*

Mentors: Tim Ophelders, Tao Hou

References

[1] CosmoSim Database. Bolshoi particle data. `https://www.cosmosim.org/datalink/bolshoi_particles/`, 2022.

[2] Georg Wilding, Keimpe Nevenzeel, Rien van de Weygaert, Gert Vegter, Pratyush Pranav, Bernard J.T. Jones, Konstantinos Efstathiou, and Job Feldbrugge. Persistent homology of the cosmic web – I. Hierarchical topology in ΛCDM cosmologies. *Monthly Notices of the Royal Astronomical Society*, 507(2):2968–2990, 2021.

2.4 Directional transforms

In directional transforms for topological data analysis, instead of taking one persistence diagram or Reeb graph, one considers an infinite family of them. In [2], they prove that an infinite set of these completely determine the input shape, so the construction is invertible. Later work considers how many directions are needed to completely determine simplicial complexes in $\mathbb{R}^d$, and proves the number necessary is exponential in the dimension [1, 3]. However, in most applications areas people will settle for a fixed number of directions for computational reasons.

Problem 5. *How good is only a small number of directions in estimating shape in practice? Are there any "nice" classes of graphs or shapes for which we can prove bounds, or can we show that in practice a small number of directions suffices for some situations, such as embedded planar graphs?*

Mentors: Tim Ophelders, Erin Chambers

References

[1] Robin Lynne Belton, Brittany Terese Fasy, Rostik Mertz, Samuel Micka, David L. Millman, Daniel Salinas, Anna Schenfisch, Jordan Schupbach, and Lucia Williams. Reconstructing embedded graphs from persistence diagrams. *Computational Geometry*, 90:101658, 2020.

[2] Justin Curry, Sayan Mukherjee, and Katharine Turner. How many directions determine a shape and other sufficiency results for two topological transforms, 2021.

[3] Brittany Terese Fasy, Samuel Micka, David L. Millman, Anna Schenfisch, and Lucia Williams. A faithful discretization of the augmented persistent homology transform, 2022.

2.5 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 6. *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 7. *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.