TDAstats: R pipeline for computing persistent homology in topological data analysis

Raoul R. Wadhwa[1]**, Drew F.K. Williamson**[2]**, Andrew Dhawan**[3]**, and Jacob G. Scott**[4]

1 Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA **2** Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA **3** Neurological Institute, Cleveland Clinic Foundation, Cleveland, OH 44195, USA **4** Department of Translational Hematology and Oncology Research, Cleveland Clinic Foundation, Cleveland, OH 44195, USA

Summary

High-dimensional datasets are becoming more common in a variety of scientific fields. Well-known examples include next-generation sequencing in biology, patient health status in medicine, and computer vision in deep learning. Dimension reduction, using methods like principal component analysis (PCA), is a common preprocessing step for such datasets. However, while dimension reduction can save computing and human resources, it comes with the cost of significant information loss. Topological data analysis (TDA) aims to analyze the "shape" of high-dimensional datasets, without dimension reduction, by extracting features that are robust to small perturbations in data. Persistent features of a dataset can be used to describe it, and to compare it to other datasets. Visualization of persistent features can be done using topological barcodes or persistence diagrams (see figure). Application of TDA methods has granted greater insight into high-dimensional data (Lakshmikanth et al. 2017); one prominent example of this is its use to characterize a clinically relevant subgroup of breast cancer patients (Nicolau, Levine, and Carlsson 2011).

The `TDAstats` R package is a comprehensive pipeline for conducting TDA. Once data is loaded into R, `TDAstats` can calculate, visualize, and conduct nonparametric statistical inference on persistent homology. The Ripser C++ library (Bauer 2015), benchmarked at approximately 40 times faster than comparable software, is wrapped using Rcpp (Eddelbuettel and Francois 2011) for efficient computation of persistent homology. `TDAstats` generates topological barcodes and persistence diagrams using the ubiquitous ggplot2 library (Wickham 2016), allowing use of ggplot2 functions to manipulate plots. This reduces the number of manual steps required to prepare publication-quality figures, thus enabling reproducible research (Sandve et al., n.d.). `TDAstats` also implements nonparametric hypothesis testing of persistent homology using a permutation test, first described by Robinson and Turner (2017). To our knowledge, `TDAstats` is the first library to implement this feature.

The primary barrier to using TDA is not mathematical comprehension. Although the algebraic topology that underlies TDA requires graduate-level study, the concepts necessary for application of TDA are far more intuitive. Rather, the barrier to entry is the lack of accessible, user-friendly software. `TDAstats` has an easy-to-use API with only 4 functions, each with only one or two intuitive parameters. Additionally, the provided vignettes cover its functionality with a comprehensive introduction and case study. Thus, even minimal
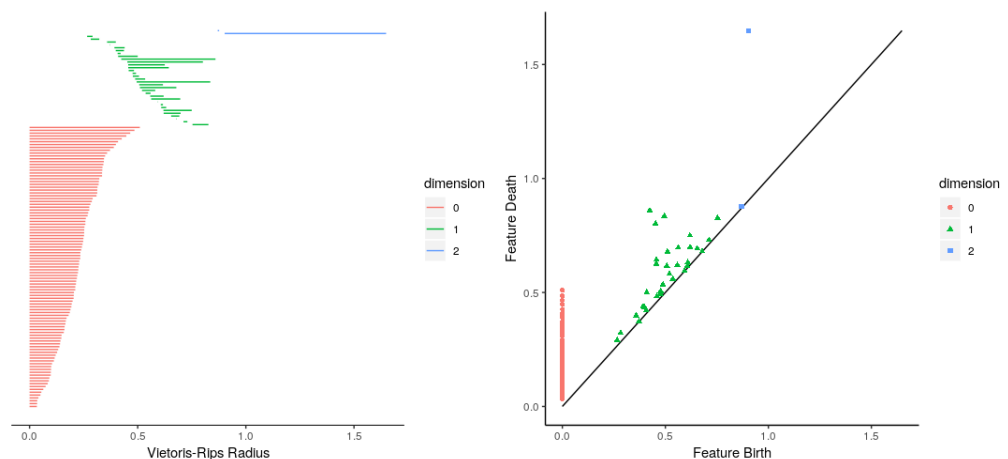
Figure 1: Topological barcode (left) and persistence diagram (right) of the sphere3d sample dataset included with `TDAstats`. The 0-cycles are colored red, the 1-cycles are colored green, and the 2-cycles are colored blue. For details on interpreting these plots, see Otter et al. (2017).

knowledge of R will be sufficient to conduct TDA. We intend to use `TDAstats` to improve digit recognition algorithms, and hope that, with its efficient implementation and user-friendly API, a far larger set of students and researchers can now apply TDA to answer research questions.

Acknowledgments

References

Bauer, Ulrich. 2015. *Ripser: A Lean C++ Code for Computation of Vietoris-Rips Persistence Barcodes.* https://github.com/Ripser/ripser.

Eddelbuettel, Dirk, and Romain Francois. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40 (8):1–18. https://doi.org/10.18637/jss.v040.i08.

Lakshmikanth, Tadepally, Axel Olin, Yang Chen, Jaromir Mikes, Erik Fredlund, Mats Remberger, Brigitta Omazic, and Petter Brodin. 2017. "Mass Cytometry and Topological Data Analysis Reveal Immune Parameters Associated with Complications After Allogeneic Stem Cell Transplantation." *Cell Reports* 20:2238–50. https://doi.org/10.1016/j.celrep.2017.08.021.

Nicolau, Monica, Arnold J. Levine, and Gunnar Carlsson. 2011. "Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival." *Proc Natl Acad Sci USA* 108 (17):7265–70. https://doi.org/10.1073/pnas.1102826108.

Otter, Nina, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. 2017. "A Roadmap for the Computation of Persistent Homology." *EPJ Data Science* 6:17. https://doi.org/10.1140/epjds/s13688-017-0109-5.

Robinson, Andrew, and Katharine Turner. 2017. "Hypothesis Testing for Topological Data Analysis." *Journal of Applied and Computational Topology* 1 (2):241–61. https://doi.org/10.1007/s41468-017-0008-7.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. n.d. "Ten Simple Rules for Reproducible Computational Research." *PLoS Computational Biology* 9 (10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.