




TDApplied: An R package for machine learning and inference with persistence diagrams

Shael Brown ¹ and Reza Farivar-Mohseni ²

¹ Department of Quantitative Life Sciences, McGill University, Montreal, Canada ² McGill Vision Research, Department of Ophthalmology, McGill University, Montreal, Canada

DOI: [10.21105/joss.06321](https://doi.org/10.21105/joss.06321)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [AHM Mahfuzur Rahman](#)

Reviewers:

- [@EduPH](#)
- [@peekxc](#)

Submitted: 25 January 2024

Published: 27 March 2024

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Topological data analysis is a collection of tools, based on the mathematical fields of topology and geometry, for finding structure in whole datasets. Its main tool, persistent homology (Edelsbrunner et al., 2000; Zomorodian & Carlsson, 2005), computes a shape descriptor of a dataset called a persistence diagram which encodes information about holes that exist in the dataset (example applications span a variety of areas, see for example Gracia-Tabuenca et al. (2020), Haim Meiom & Bobrowski (2022), and Krishnapriyan (2021)). These types of features cannot be identified by other methods, making persistence diagrams a unique and valuable data science object for studying and comparing datasets. The two most popular data science tools for analyzing multiple objects are machine learning and inference, but to date there has been no open source implementation of published methods for machine learning and inference of persistence diagrams.

Statement of need

TDApplied is the first R package for machine learning and inference of persistence diagrams, building on the main R packages for the calculation of persistence diagrams TDA (Fasy et al., 2021) and TDAstats (R. Wadhwa et al., 2019; R. R. Wadhwa et al., 2018) and publications of applied analysis methods for persistence diagrams (Le & Yamada, 2018; Robinson & Turner, 2017). TDApplied is intended to be used by academic researchers and industry professionals wanting to integrate persistence diagrams into their analysis workflows. An example TDApplied workflow, in which the topological differences between three datasets are visualized in 2D using multidimensional scaling (MDS) (Cox & Cox, 2008), is visualized in [Figure 1](#):

Visualizing topological differences between multiple datasets

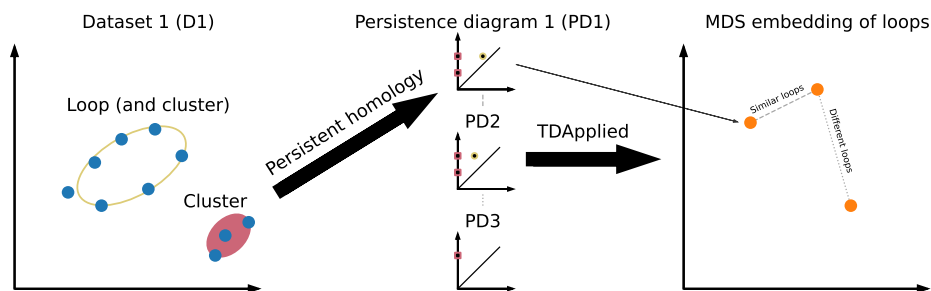


Figure 1: An example TDApplied workflow. A dataset (D1, left) contains one loop (yellow) and two clusters (the loop forms one cluster and the three points on the bottom are another cluster, and clusters are denoted by the color red). These topological features are captured with persistent homology in a persistence diagram PD1 (middle top), and two other data sets, D2 and D3 (not shown), have their persistence diagrams, PD2 and PD3, computed (middle center and middle bottom). PD1 and PD2 are not very topologically different in terms of their loops, with both containing a loop with similar birth and death values, and this is represented by a dashed-line relationship. On the other hand, PD2 and PD3 are topologically different in terms of their loops because PD3 does not contain a loop, and this is represented by a dotted-line relationship. TDApplied can quantify these topological differences and use MDS to project the persistence diagrams into three points in a 2D embedding space (right) where interpoint distances reflect the topological differences between the persistence diagrams.

The TDApplied package is built on three main pillars:

1. User-friendly – internal preprocessing of persistence diagrams that would normally be left to R users to figure out ad hoc, and functions designed to easily flow from input diagrams to output metrics.
2. Efficient – parallelization, C code, computational tricks and storage of reusable and cumbersome calculations significantly increases the feasibility of topological analyses (compared to existing R packages).
3. Flexible – ability to interface with other data science packages to create personalized analyses.

TDApplied has already been featured in a [conference workshop](#) and a [conference tutorial](#), utilized in a journal publication ([Singh et al., 2023](#)) and downloaded over 4400 times. Therefore, we propose TDApplied as a user-friendly, efficient and flexible R package for the analysis of multiple datasets using machine learning and inference via topological data analysis.

Project Management

Installation and availability: TDApplied can be installed directly from CRAN using the command `install.packages("TDApplied")`, or from GitHub using the `devtools` package ([Wickham et al., 2021](#)). TDApplied is distributed under the GPL-3 license.

Code quality: Code has been tested using the `testthat` package ([Wickham, 2011](#)), with 91.45% coverage of R code when not skipping tests involving Python code (or 88.44% coverage when skipping the Python tests).

Documentation: TDApplied contains five main vignettes:

1. “TDApplied Theory and Practice” provides example function usage on simulated data as well as mathematical background and intuition,
2. “Human Connectome Project Analysis” demonstrates an applied example analysis of neurological data,

3. “Benchmarking and Speedups” outlines the package’s optimization strategies and highlights performance gains compared to other packages,
4. “Personalized Analyses with TDApplied” demonstrates how to interface TDApplied with other data science packages, and
5. “Comparing Distance Calculations” accounts for differences in computed distance values between persistence diagrams across comparable packages.

Acknowledgements

We acknowledge funding from the CIHR 2016 grant for cortical mechanisms of 3-D scene and object recognition in the primate brain.

References

- Cox, M. A. A., & Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization* (pp. 315–347). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_14
- Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2000). Topological persistence and simplification. *Discrete & Computational Geometry*, 28, 511–533. <https://doi.org/10.1007/s00454-002-2885-2>
- Fasy, B. T., Kim, J., Lecci, F., Maria, C., Millman, D. L., & Rouvreau, V. (2021). *TDA: Statistical tools for topological data analysis*. <https://CRAN.R-project.org/package=TDA>
- Gracia-Tabuenca, Z., Diaz-Patino, J. C., Arelio, I., & Alcauter, S. (2020). Topological data analysis reveals robust alterations in the whole-brain and frontal lobe functional connectomes in attention-deficit/hyperactivity disorder. *Eneuro*. <https://doi.org/10.1523/eneuro.0543-19.2020>
- Haim Meir, S., & Bobrowski, O. (2022). Unsupervised geometric and topological approaches for cross-lingual sentence representation and comparison. *Proceedings of the 7th Workshop on Representation Learning for NLP*, 173–183. <https://doi.org/10.18653/v1/2022.repl4nlp-1.18>
- Krishnapriyan, A. S. et al. (2021). Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Nature Scientific Report*, 11. <https://doi.org/10.1038/s41598-021-88027-8>
- Le, T., & Yamada, M. (2018). Persistence Fisher kernel: A Riemannian manifold kernel for persistence diagrams. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. <https://doi.org/10.48550/arXiv.1802.03569>
- Robinson, A., & Turner, K. (2017). Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, 1. <https://doi.org/10.1007/s41468-017-0008-7>
- Singh, Y., Farrelly, C. M., Hathaway, Q. A., Leiner, T., Jagtap, J., Carlsson, G. E., & Erickson, B. J. (2023). Topological data analysis in medical imaging: Current state of the art. *Insights into Imaging*, 14(1), 58. <https://doi.org/10.1186/s13244-023-01413-w>
- Wadhwa, R. R., Williamson, D. F. K., Dhawan, A., & Scott, J. G. (2018). TDAstats: R pipeline for computing persistent homology in topological data analysis. *Journal of Open Source Software*, 3(28), 860. <https://doi.org/10.21105/joss.00860>
- Wadhwa, R., Dhawan, A., Williamson, D., & Scott, J. (2019). *TDAstats: Pipeline for topological data analysis*. <https://github.com/rrrlw/TDAstats>

- Wickham, H. (2011). testthat: Get started with testing. *The R Journal*, 3, 5–10. <https://doi.org/10.32614/rj-2011-002>
- Wickham, H., Hester, J., Chang, W., & Bryan, J. (2011). devtools: Tools to make developing R packages easier. <https://CRAN.R-project.org/package=devtools>
- Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete and Computational Geometry*, 33, 249–274. <https://doi.org/10.1007/s00454-004-1146-y>