Modular, interoperable, and extensible topological data analysis in R.

Jason Cory Brunson and Aymeric Stamm

2024-04-01

Signatories

Project team

Jason Cory Brunson, PhD (Project Co-lead): Dr. Brunson (Assistant Professor, University of Florida) is a mathematician by training with a long-running interest in topological data analysis (TDA). He has authored or contributed to several TDA packages on CRAN, including {TDAstats}, {ripserr}, {simplextree}, {tdaunif}. He also has two feature-complete packages in preparation, a {ggplot2} extension {ggtda} for topological data and an R interface {plt} to the Persistence Landscapes Toolbox.

Aymeric Stamm, PhD (Project Co-lead): Dr. Stamm (Research Engineer, CNRS, France) is a mathematical engineer by training with specialisation in statistics. He is in particular the author and maintainer of the {rgudhi} package which ports the GUDHI library into R and of the {flipr} package for permutation-based inference. He has extensive expertise on efficient programming in R including optimisation of R code, use of C/C++ code through {Rcpp} and parallelisation of either R or C++ code.

Paul Rosen, PhD (Advisory Board Member): Dr. Rosen (Associate Professor, University of Utah) is a computer scientist who focuses on geometric and topological approaches in scientific and information visualization and has authored several computational tools for TDA.

Bertrand Michel, PhD (Advisory Board Member): Dr. Michel (Professeur, Ecole Centrale Nantes, France) is a mathematician who focuses on statistical methods for geometric and topological inference, TDA, model selection and high dimensional statistics. He is a collaborator and member of the advisory board of the GUDHI Python and C++ library for TDA.

Contributors

Several related packages, dovolped in collaboration with Raoul Wadhwa, Matt Piekenbrock, Yara Skaf, James Otto, and others are published on CRAN, installable from GitHub, or in development.

Consulted

We have solicited feedback on drafts of this proposal from members of (1) the TDA Discussion Group at the University of Florida (UF)[1], (2) the Laboratory for Systems Medicine at the University of Florida[2], and (3) presenters at UFTDA2024[3], a conference organized by the Bubenik lab and hosted at UF.

The Problem

R is one of the most widely used programming languages in statistics. It is used by professionals ranging from data analysts to educators to research statisticians, to implement newly developed methods and facilitate hands-on learning as well as to conduct routine analyses. Topological data analysis is an emerging and

[1] https://people.clas.ufl.edu/nikola/topological-data-analysis-discussion-group/

[2] https://systemsmedicine.pulmonary.medicine.ufl.edu/

[3] https://people.clas.ufl.edu/peterbubenik/uftda2024/

heterogeneous area of statistical research that is grounded in the mathematical discipline of topology. It intersects heavily with exploratory analysis and visualization, statistical inference, and machine learning. It is therefore important to the currency of mainstream statistical work and to ongoing methods development for R users to have access to comprehensive and reliable TDA tools.

At present, this is not the case. Feedback from colleagues has been consistent on the following points. Regarding TDA software in general:

1. Most techniques used in TDA are not easily accessible or well documented and require significant training or experimentation to use.
2. Most practitioners implement at least some tools described in published studies from scratch.

Regarding solutions to the problems motivating this proposal:

3. No software resource provides a comprehensive set of vectorization methods for persistence data.
4. Most hypothesis-testing work is conducted using original *ad hoc* code.
5. Practitioners frequently use (various implementations or ports of) Ripser in longer computational pipelines that would otherwise be impractical.

This feedback suggests that practitioners face challenges when using TDA software that, though they are overcome by specialists, are likely to be daunting to non-specialist researchers and analysts.

Published R packages for TDA fall into three categories: First, {TDA} and {rgudhi} provide interfaces to comprehensive libraries in lower-level languages, including GUDHI, Dionysus, and PHAT. While they provide essential tools to users—{TDA} is the most frequently downloaded of the packages surveyed here—they are not designed to integrate with common tools and workflows outside TDA. They are also made fragile by the need to adapt to upgrades both in the source code language and in R. Second, {TDAstats}, {TDAkit}, {TDApplied}, and {GSSTDA} provide bespoke toolkits for statistical inference, machine learning, and survival analysis. These packages tend to be designed for specialists rather than general users, self-contained rather than modular, syntactically inharmonious with common workflows, and reliant on libraries written in other languages. These limitations have likely hindered the adoption of TDA approaches by non-specialists; for example, an Rseek search for "topological data analysis" returns a handful of commentaries, tutorials, and workshop notes, most of which use only one package and all of which are several years old.[4]

A third category of TDA packages fit a different profile, in that each package performs a narrow task: {simplextree} provides the simplex tree data structure, {interplex} converts between this and other data structures for simplicial complexes, {ripserr} interfaces to the Ripser algorithm to compute persistent homology, and {tdaunif} provides uniform samplers for topologically interesting spaces. These packages have been developed by Dr. Brunson and colleagues with the goal of building a general-purpose, native, modular, interoperable, and extensible R package collection for TDA. They are not as comprehensive as the first category, and they are not yet as interoperable and extensible as envisioned. In order to meet the needs of non-specialists, the collection will need to resolve its own incompatibilities and be integrated into more widely used data analysis tools.

The proposal

Our **goal** is to seamlessly integrate popular techniques from topological data analysis (TDA) into common statistical workflows in R. The **expected benefit** is that these extensions will be more widely used by non-specialist researchers and analysts, which will create sufficient awareness and interest in the community to extend the individual packages and the collection.

[4] https://blog.revolutionanalytics.com/2014/01/topological-data-analysis-with-r.html, https://rviews.rstudio.com/2018/11/14/a-mathematician-s-perspective-on-topological-data-analysis-and-r/, https://calebntorres.github.io/pdf/TDA.pdf, https://people.clas.ufl.edu/peterbubenik/files/tda_r_workshop.pdf, https://gist.github.com/rrrlw/2fd22a834a883cb66454b1dabab9fdcb

Overview

Tidymodels (Kuhn and Wickham 2020) provides a complete toolkit for common machine learning (ML) tasks. ML relies heavily on vectorizations, and the last decade of work in TDA has produced several for topological data. Our **first aim** is to make these available in a Tidymodels-compatible R package.

TDA relies heavily on statistical inference. A variety of hypothesis tests have been proposed in the research literature and have idiosyncratic implementations. Our **second aim** is to provide a Tidymodels-compatble package for permutation-based hypothesis testing with topological data.

The primary workhorse for TDA is Ripser, which efficiently computes Vietoris–Rips filtrations. The engine used in the lightweight wrapper package {ripserr} is missing key features, so our **third aim** is to upgrade it.

Detail

Aim 1: Publish a {recipes} extension for ML vectorizations based on persistent homology.

The new package would provide `step_*()` functions for several topological transformations and document the process for contributing additional steps. From among many proposed transformations (Fasy et al. 2020; Pun, Lee, and Xia 2022; Ali et al. 2022), we intend to include the following in a first release based on their relative simplicity, time on CRAN, frequency of use, and expected implementation cost: persistence statistics, Betti curve, lifespan curve (to be implemented anew), persistence landscape (to rely on the forthcoming {plt}), and persistent image (to be adapted from the original Matlab implementation [5]). Dr. Brunson has experience writing an unpublished but installable {recipes} extension for association rules.[6]

Aim 2: Publish a {flipr} extension for permutation-based statistical inference on topological data.

Dr. Stamm leads a coordinated package collection for permutation-based inference, with the published {flipr} package at its center. This package features (i) implementations of classic permutation schemes that are agnostic to the type of input and (ii) the central `PlausibilityFunction` R6 class for seamless hypothesis testing and confidence estimation. Several extensions are published or in development for specific data types, including {nevada} for network-valued data, {fdatest} for functional data, and unpublished extensions for scalar- and vector-valued data. The new package will extend methods for two-sample testing, ANOVA, hypothesis testing, and confidence estimation to data types arising from TDA. It will inherit non-parametric combination (NPC) (Pesarin and Salmaso 2010), which enables the use of several test statistics in a single hypothesis test, making the combination sensitive to different aspects of the compared underlying distributions.

Aim 3: Refactor {ripserr} with a current implementation of Ripser and connect additional options to R.

The base implementation of Ripser is written in C++, and {ripserr} provides integration via {Rcpp}. We therefore anticipate that this update will be a straightforward, though not trivial, exercise in C++/R integration. In particular, it will provide options to retrieve representative cycles and cocycles for topological features, which are essential for many practical applications.

Project plan

Start-up phase

Development will take place publicly on two GitHub accounts: TDAverse, co-owned by Dr. Brunson (Aims 1 and 3), and LMJL-Alea, co-owned by Dr. Stamm (Aim 2). Both PIs will have write access to the repositories for all project packages. All project packages will be licensed compatibly with R and Tidyverse/Tidymodels (MIT or GPL >= 2). The PIs will provide the R Consortium with monthly updates based on progress toward the deliverables below.

[5] https://github.com/CSU-TDA/PersistenceImages

[6] https://github.com/corybrunson/arulesteps

Technical delivery

Aim 1: {recipes} extension for topological data transformations (<= 16wk)

1. Assembly of external engines or original implementations of selected transformations. (<= 6wk)
2. Template {recipes} extension, to ensure that infrastructural checks are satisfied. (<= 2wk)
3. Complete implementation of a single transformation, with documentation of all necessary steps (including unit tests, documentation, and examples). (<= 2wk)
4. Implementations of remaining selected transformations. (<= 4wk)
5. Vignette illustrating complete modeling workflow using at least two transformations. (<= 2wk)

Aim 2: {flipr} extension for topological data types (<= 10wk)

1. Review and selection of a first set of metric spaces (representation + metric) into which the persistence data is to be analysed and implementation of such spaces. This will probably focus on vectorized representations, which would be assembled jointly with the {recipes} extension from Aim 1. (<= 2wk)
2. Implementations of a variety of test statistics that focus on comparing different features between distributions, either exploiting accessibility to Frechet means and variances when the metric space allows for it or resorting to statistics based on inter-point distances. (<= 2wk)
3. Implementations of confidence bands for functional representations of persistance data based on above implemented test. (<= 4wk)
4. Writing up a vignette illustrating new functionalities. (<= 2 wk)

Aim 3: {ripserr} upgrade (<= 8wk)

1. Selection and reconciliation of current Ripser implementations in C++. (<= 4wk)
2. Integration into package using {Rcpp}. (<= 2wk)
3. Updated vignette illustrating new functionality. (<= 2wk)

Other aspects

Each package release will be announced on social media platforms using the `#rstats` tag and on blogs cross-posted to R-bloggers. We will submit abstracts for each aim to relevant conferences (e.g., useR!, `posit::conf()`, the Joint Statistical Meetings). We will also propose in-person workshops co-located with local or regional conferences. Finally, we plan to submit at least two journal articles: one on a handful of packages that integrate with the Tidyverse (including the {recipes} extension and the nearly complete {ggplot2} extension {ggtda}), another on the {flipr} collection including its TDA extension.

Requirements

People

The PIs (Drs. Brunson and Stamm) will lead development on the three aims. Both PIs routinely recruit undergraduate research assistants. As interested students are found, this grant will support their contributions.

Processes

For first releases, no new processes are needed: Packages will be accessible and contributable to by way of detailed instructions based on those of other packages in the Tidyverse and the PIs' portfolios. Future releases will rely on user community input to identify bugs and prioritize features. While no plan is proposed, the PIs are prepared to discuss handoffs to interested user–contributors.

Tools & Tech

Development will depend only on the continued upkeep of dependencies, in particular the Tidymodels collection and {Rcpp}, and on the availability of a collaborative version control service like GitHub. Component tools may require original implementations in R but no original theory.

Funding

The timeline above envisions both PIs contributing 1/6 FTE during an 8-month award period. At a mean salary of about $80,000/year, this amounts to roughly $18,000 total. Up to 80% of this total may be redirected to students through mentored projects. A smaller award amount would support Aims 1 and 2 to the extent possible, with Aim 3 as a stretch goal.

Summary

Time is expected to be spread across tasks as follows: development and documentation (75%), publication and promotion (15%), coordination and mentorship (10%).

Success

Definition of done

Our aspirations for this project are a more stable TDA ecosystem in R that supports greater efficiency of TDA research and analysis by R users.

Measuring success

Our benchmarks for project completion are twofold: (1) publication of packages on CRAN, and (2) complete workflows, published in vignettes, blog posts, or software journal articles, that use these tools to perform typical analysis and modeling tasks in practical settings, including using real-world data.

After completion, we will follow several indicators to judge success and prompt necessary follow-up work, including monthly download rates, questions and problems raised on Stack Exchange (in addition to issues raised on the package repositories), blog posts by users, citations in published research, and user contributions to code base. That last would be a very strong indicator of success.

Future work

Much will remain after these aims to make this modular R package ecosystem for topological data analysis more comprehensive. Immediate next goals include (1) a new lightweight package of basic tools for handling persistence data, such as an efficient data structure and common distance measures; (2) a revived {Mapper} package for exploratory TDA (Singh, Mémoli, and Carlsson 2007), from which additional multi-purpose low-level packages may be spun off; and (3) tools for working with Reeb graphs, such as the `ReebGraphPairing` Java package [7] to compute extended homology.

Key risks

We foresee two primary risks to this proposal: Before publication, the greatest risk is that structures and procedures used in Matlab and C++ may not translate easily to R. Alternative solutions are to implement directly from theory (persistent images) and to include separate implementations for different tasks (Ripser). After publication, the greatest risk is lack of community uptake. It may be that there is and will be less appetite than we expect for more customizable tools outside the specialist community. To reduce this risk, we focus this proposal on the most common yet least aavailable tools, based on conversations with colleagues.

References

Ali, Dashti, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, and Manuel Soriano-Trigueros. 2022. "A Survey of Vectorization Methods in Topological Data Analysis." arXiv. https://doi.org/10.48550/arXiv.2212.09703.

[7] https://github.com/USFDataVisualization/ReebGraphPairing

Fasy, Brittany, Yu Qin, Brian Summa, and Carola Wenk. 2020. “Comparing Distance Metrics on Vectorized Persistence Summaries.” In *NeurIPS 2020 Workshop on Topological Data Analysis and Beyond*.

Kuhn, Max, and Hadley Wickham. 2020. “Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.”

Pesarin, Fortunato, and Luigi Salmaso. 2010. “The Permutation Testing Approach: A Review.” *Statistica* 70 (4): 481–509. https://doi.org/10.6092/issn.1973-2201/3599.

Pun, Chi Seng, Si Xian Lee, and Kelin Xia. 2022. “Persistent-Homology-Based Machine Learning: A Survey and a Comparative Study.” *Artificial Intelligence Review* 55 (7): 5169–5213. https://doi.org/10.1007/s10462-022-10146-z.

Singh, Gurjeet, Facundo Mémoli, and Gunner Carlsson. 2007. “Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.” In *Eurographics Symposium on Point-Based Graphics*, edited by M. Botsch, R. Pajarola, B. Chen, and M. Zwicker. The Eurographics Association. https://doi.org/10.2312/SPBG/SPBG07/091-100.