Postdoc Journal
Vol. 4, No. 1, January 2016

Journal of Postdoctoral Research
www.postdocjournal.com

Statistical Techniques in TDA with Applications to Real Data

Brittany Terese Fasy, Ph.D.
Postdoc Institution: Department of Computer Science, Tulane University
Current Institution: Computer Science Department, Montana State University
Email: brittany@fasy.us

Abstract Persistent homology is a widely used tool in Topological Data Analysis that encodes multi-scale topological information as a multi-set of points in the plane, called a persistence diagram. Each of these persistence points is associated with a lifetime (or persistence). Features with short lifetimes are informally considered to be *topological noise*, and those with a long lifetime are considered to be *topological signal*. We bring some statistical ideas to persistent homology in order to derive confidence sets that allow us to separate topological signal from topological noise. We also apply statistical theory to other topological descriptors such as the persistence landscape or silhouette, rather than working with the original diagrams or data sets. We motivate this work with three applications.

1 Introduction

Understanding and comparing data sets is a challenge present across various disciplines. Succinctly representing a data set as a structure known as a persistence diagram allows us to easily visualize and compare the (topological) features present in a data set. This is an example of *topological* data analysis (TDA). In this paper, we focus on the integration of statistical methods and TDA.

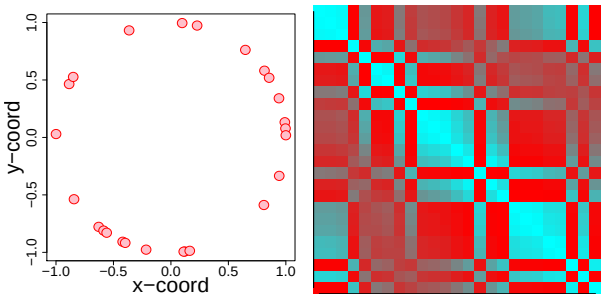

Figure 1: Left: 25 Points sampled from the uniform distribution over the uniform circle. Right: The corresponding pairwise distance matrix (cyan represents small distance).

2 Topological Data Analysis

TDA is a field that stems from Morse Theory [25]. As its own research field, TDA developed in the late 90s; see [18, 23]. Within the past five years, TDA has really exploded as a research field, including mathematicians, computer scientists, statisticians, as well as domain experts in varying fields from astronomy [31] to cancer research [28] to natural language processing [33].

When faced with a data analysis task in any field, we are often interested in summarizing and comparing data sets. Often, working with the data directly is cumbersome. This is where persistent homology comes into play. Persistent homology takes a data set describes it succinctly with a finite set of points in the plane, called a persistence diagram. Each point in the persistence diagram represents a topological feature that is present for some view of the data set. The views of interest depend on the application, but are often a range of either times or distances. The point encodes the smallest (*birth*) time for which the feature is present and the largest (*death*) time for the feature.

For example, consider a set of points sampled from a circle, as shown in Figure 1. We compute the pairwise distances between every

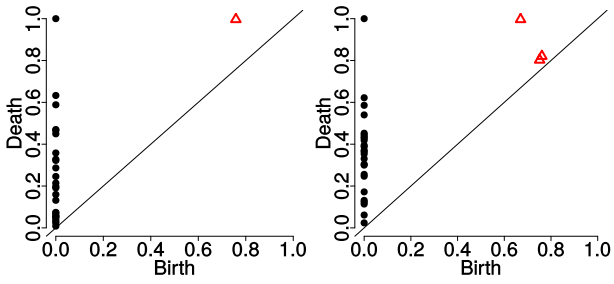

Figure 2: Left: Persistence diagram for the VR filtration of the example of Figure 1. Right: Diagram obtained when data points are perturbed.

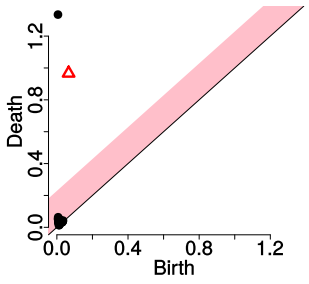

Figure 3: A persistence diagram with a confidence band (pink). Notice that there are two statistically significant persistence points.

pair of points to obtain the (symmetric) pairwise distance matrix M. Using M, we construct the Vietoris-Rips (VR) filtration. The VR filtration is an indexed sequence of nested simplicial complexes which starts with the complex consisting of n vertices, one corresponding to each data point, and ends with the complete complex on these n vertices. Connected components and loops are tracked as the index of the filtration increases, giving rise to the persistence diagram; see Figure 2 (left).

The benefit of considering the persistence diagrams, as opposed to the original raw data set, is that (i) it captures patterns such as connected components and loops, and (ii) we can compute a meaningful distances between persistence diagrams, such as the Wasserstein or Bottleneck distances [16, 17]. Moreover, persistence diagrams are stable [8, 9, 14, 15]. That is, small changes in the filtration (e.g., changes induced by small changes in the data set) will result in small changes in the persistence diagram. Figure 2 (right) is a diagram computed after slightly perturbing the coordinates of the data points shown in Figure 1. The black dots, which correspond to components, have moved slightly up or down. What is most notable, however, is the fact that two new loops (denoted by the red triangles) have emerged. We observe that these persistence points are close to the diagonal, and hence have a very short persistence (interval of existence in the filtration). Leveraging the stability results mentioned above, we can use the bottleneck distance in order to compare and to cluster data sets, as we will explore in the next section.

3 Integrating Statistics into TDA

Mileyko et al. studied the space of persistence diagrams under the Wasserstein metric, observing that the space is complete and separable [24]. Moreover, finding the (Fréchet) average of persistence diagrams was studied in [26, 30]. However, as was observed, the average diagram is not unique, so concisely representing the means can become problematic. To overcome this issue, one could instead compute persistence landscapes and silhouettes, which do have a unique mean [7, 13]. Taking the mean of landscapes and silhouettes allows for statistical analysis, but the trade-off is that it is more obscure to interpret than a persistence diagram.

Fasy et al. [21] defined a confidence set for a persistence diagram. In particular, if $\widehat{D}$ is an estimate of the unknown diagram D, we say that $\widehat{D}$ and a distance δ define a *confidence set* for D if:

$$\mathbb{P}\left(d(D,\widehat{D}) \leq \delta\right) \geq 1-\alpha,$$

where $d(\cdot,\cdot)$ is an appropriate distance measure between persistence diagrams and α is a confidence level. In words, any point farther than δ from the diagonal is statistically significant. We

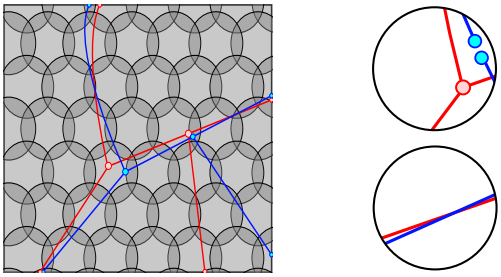

Figure 4: Left: Covering the domain with overlapping local neighborhoods. Right: Two local neighborhoods illustrating a large local distance (top) and a small local distance (bottom).

illustrate this in Figure 3, where we see one significant cycle and one significant component. Further analysis of these confidence sets, of convergence, and of limiting distributions for persistence diagrams and landscapes was studied in [10, 11, 12, 13]

The techniques mentioned above have been made accessible in the R package TDA[1], released on CRAN in 2014; see [20] for a user guide.

4 Applications

TDA has the potential to be a powerful analysis tool in many domains. In particular, we highlight three application areas to demonstrate the breadth of TDA.

4.1 Road Network Analysis

Today, having your phone or other device record your GPS location has become commonplace: runners and bikers use it to track their progress, and companies such as Foursquare collect this data in order to know how to best advertise to users of their apps. Considering the former scenario, trajectories in the woods and parks often do not have recorded paths, so it is not possible to (meaningfully) snap the paths to a road (or trail) network. However, one can reconstruct the road network from the GPS trajectories [4]. Most recently, using kernel density estimates and topological methods has been proposed [1, 32].

The question remains: how accurate are these reconstructed road networks? To answer this question, we turn to a concept called persistent local homology [5, 6, 22]. In particular, for a road network embedded in a compact domain X, we cover X with a set of neighborhoods centered on a regular lattice; see Figure 4. Then, we compare the local structures as witnessed by these neighborhoods using persistent local homology and aggregate these local distances [2]. Other road network comparison algorithms and heuristics do exist; see [3]. However, a full discussion of these is beyond the scope of the current paper.

4.2 Distribution of Galaxies

That the distribution of galaxies follows a Voronoi diagram-like pattern comprising of clusters, filaments, and sheets is a well-established assumption in astronomy. These sheets enclose *voids*, which many astronomers are interested in better understanding [27, 29]. With the release of the Sloan Digital Sky Survey (SDSS)[2], an inventory of astronomical objects (galaxies, stars, quasars) is now publicly and freely available.

One of the big open questions in astronomy is: do the observations that are collected match the models that have been developed? To answer this question, we look at collection of *cubes* of space both from observations (SDSS data set) as well as from the models. We then ask: are these two sets of data cubes collected from the same distribution? Or, is there a critical flaw or bias in the models? The TopStat[3] research group is currently developing the statistical theory necessary to justify the use of various hypothesis tests in TDA, focusing on this application of comparing models and observations of the distribution of celestial objects.

[1] `https://cran.r-project.org/web/packages/TDA/index.html`

[2] `http://www.sdss.org/`

[3] `http://www.stat.cmu.edu/topstat/`

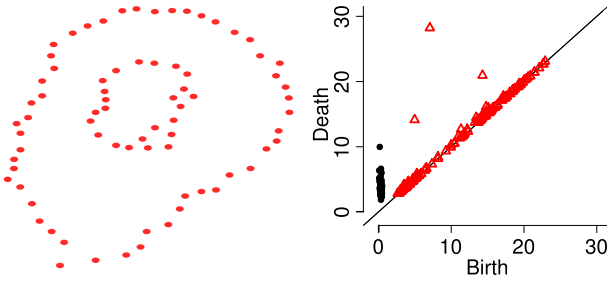

Figure 5: Left: Nuclei forming a telescoping gland pattern, indicative of Gleason pattern three. Right: The corresponding persistence diagram, showing multiple loops are present.

4.3 Prostate Cancer Biopsy

Prostate cancer is one of the most common forms of cancer worldwide. The diagnosis of prostate cancer often involves taking a biopsy and assigning it a primary and a secondary Gleason grade. The Gleason grade ranges from 1 (normal-looking) to 5 (extremely irregular and carcinogenic). Pathologists learn how to grade biopsies through a book with many illustrations of different grades [19], as well as through practice. As a result, the grading between pathologists is not always consistent. As part of a recently funded NSF-NIH planning grant, we are working on using persistence diagrams and other topological descriptors to quantify the information in prostate cancer biopsy slides. For example, in Figure 5, we see a gland and the corresponding persistence diagram. The gland shown is a representative gland of Gleason grade three, and illustrates the telescoping glandular pattern (gland within gland).

5 Discussion

This paper highlighted a few high-level results in topological data analysis, as well as to present three examples demonstrating the breadth of TDA. A more in-depth coverage of the ideas presented here can be found in the references provided. Research in TDA is currently rapidly expanding, as illustrated by the fact that the field has grown from a couple dozen to a couple hundred researchers in just a few years. As the field continues to develop, it will be the close relationship among computer scientists, mathematicians, statisticians, and field experts that will drive the research forward.

Acknowledgements. The research referenced in this paper is a result of several collaborations. The author would like to thank the following collaborators: Mahmuda Ahmed, J. Quincy Brown, Jessi Cisewski, Maia Grudzien, Jisu Kim, Pete Lawson, Fabrizio Lecci, Christopher Miller, David L. Millman, Sawyer Payne, Alessandro Rinaldo, Ryan Thompson, Larry Wasserman, Yusu Wang, and Carola Wenk. Special thanks goes to Carola Wenk and Larry Wasserman for nominating her for this award. Some of the research presented in this paper is supported by the National Science Foundation and the National Institutes of Health under Award No. 1557716.

References

[1] Ahmed, M., Fasy, B. T., Gibson, M., and Wenk, C. Choosing thresholds for density-based map construction algorithms. In *SIGSPATIAL* (Nov. 2015), ACM.

[2] Ahmed, M., Fasy, B. T., and Wenk, C. Local persistent homology based distance between maps. In *SIGSPATIAL* (Nov. 2014), ACM.

[3] Ahmed, M., Fasy, B. T., and Wenk, C. New techniques in road network comparison. In *Grace Hopper Celebr. Women Comput.* (Oct. 2014). Online proceedings.

[4] Ahmed, M., Karagiorgou, S., Pfoser, D., and Wenk, C. A comparison and evaluation of map construction algorithms, 2014. arXiv:1402.5138.

[5] Bendich, P., Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Morozov, D. Inferring local homology from sampled stratified spaces. *Proc. FOCS* (2007), 536 – 546.

[6] Bendich, P., Wang, B., and Mukherjee, S. Local homology transfer and stratification learning. *ACM-SIAM Symp. Discrete Algorithms* (2012).

[7] Bubenik, P. Statistical topology using persistence landscapes. *JMLR* (Jan. 2015), 77–102. Also available at arXiv:1207.6437.

[8] Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L. J., and Oudot, S. Y. Proximity of persistence modules and their diagrams. In *Proc. of the* 25^{th} *Annu. Symp. Comput. Geom.* (New York, Jun. 2009), ACM. Symposium held in Aarhus, Denmark.

[9] Chazal, F., de Silva, V., Glisse, M., and Oudot, S. The structure and stability of persistence modules, 2012.

[10] Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems 20*, 6 (2013), 96–105. Also available at arXiv:1311.0376.

[11] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Robust topological inference: Distance-to-a-measure and kernel distance, 2014. In submission. Preprint available at arXiv:1412.7197.

[12] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Subsampling methods for persistent homology. In *Proc. ICML* (July 2015). Preprint available at arXiv:1406.1901.

[13] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. Stochastic convergence of persistence landscapes and silhouettes. In *Proc. of the* 30^{th} *Annu. Symp. Comput. Geom.* (Jun. 2014).

[14] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. *Discrete Comput. Geom. 37*, 1 (2007), 103–120.

[15] Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. Lipschitz functions have l_p-stable persistence. *Found. Comput. Math. 10*, 2 (Apr. 2010), 127–139.

[16] Edelsbrunner, H., and Harer, J. Persistent homology–a survey. In *Surveys on Discrete and Computational Geometry–Twenty Years Later*, vol. 453 of *Contemp. Math.* Amer. Math. Soc., 2006, pp. 257–282. AMS–IMS–SIAM Joint Summer Research Conference at Snowbird, UT.

[17] Edelsbrunner, H., and Harer, J. *Computational Topology. An Introduction.* Amer. Math. Soc., Providence, RI, 2010.

[18] Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. *Discrete Comput. Geom. 28*, 4 (Jul. 2002), 511–533.

[19] Epstein, J. I. *The Gleason Grading System: A Complete Guide for Pathologists and Clinicians.* Lippincott Williams & Wilkins, Philadelphia, PA, 2013.

[20] Fasy, B. T., Kim, J., Lecci, F., and Maria, C. Introduction to the R package TDA. In submission. Preprint available at ArXiv:1411.1830.

[21] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *Annals of Statistics 42*,

6 (2014), 2301–39. Preprint available at ArXiv:1303.7117.

[22] Fasy, B. T., and Wang, B. Exploring persistent local homology in topological data analysis. In submission.

[23] Frosini, P. Discrete computation of size functions. *Journal of Combinatorics, Information and System Sciences 17*, 3-4 (1992), 232–250.

[24] Mileyko, Y., Mukherjee, S., and Harer, J. Probability measures on the space of persistence diagrams. *Inverse Problems 27*, 12 (2011), 124007.

[25] Milnor, J. *Morse Theory*. No. 51 in Ann. of Math. Stud. Princeton Univ. P., Princeton, 1963.

[26] Munch, E., Bendich, P., Turner, K., Mukherjee, S., Mattingly, J., and Harer, J. Probabilistic Fréchet means and statistics on vineyards, 2013. arXiv 1307.6530.

[27] Neyrinck, M. C. Zobov: A parameter-free void-finding algorithm. *Monthly Not. of the Roy. Astron. Soc. 386*, 4 (2008), 2101–2109.

[28] Nicolau, M., Levine, A. J., and Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences 108*, 17 (2011), 7265–7270.

[29] Platen, E., Van De Weygaert, R., and Jones, B. J. A cosmic watershed: the wvf void detection technique. *Monthly Not. of the Roy. Astron. Soc. 380*, 2 (2007), 551–570.

[30] Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry 52*, 1 (2014), 44–70.

[31] van de Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B. J. T., Pranav, P., Park, C., Hellwing, W. A., Eldering, B., Kruithof, N., Bos, E. G. P., Hidding, J., Feldbrugge, J., ten Have, E., van Engelen, M., Caroli, M., and Teillaud, M. Alpha, betti and the megaparsec universe: On the topology of the cosmic web. In *Trans. Comput. Sci. XIV*. Springer, 2011, pp. 60–101.

[32] Wang, S., Wang, Y., and Li, Y. Efficient map reconstruction and augmentation via topological methods. In *SIGSPATIAL* (Nov. 2015), ACM.

[33] Zhu, X. Persistent homology: An introduction and a new text representation for natural language processing. In *Proc. 23rd Int. Joint Conf. AI* (2013), AAAI Press, pp. 1953–9.