1 List of works

1.1 Introduction and Summary

1. [1] Roadmap for computation of PH, 2017
   Thorough review of computation and some theory of PH.

1.2 Original work

1.2.1 TDA

1. [2] Foundation for TDA
   Discusses why topology and functoriality are essential for data analysis.
2. [3] Mapper Algorithm
3. [4] The Natural Image paper.

1.2.2 Persistent Homology

1. [5] Discrete Morse theory
2. [6] Statistical approach to persistent homology

1.2.3 Computation

1. [7] Computation of PH for cubical data.
   Found in [1]
2. [8] Computational homology
   talked about cubical complexes and other complexes
3. [9] Javaplex
4. [10] Intro to R TDA
5. [11] Intro of persistence landscapes w/ algorithms.

1.3 Applications

1.3.1 Benchmarking

1.

2 Current Challenges

2.1 Statistical Interpretation

Two major challenges are described in [1].

1. Quantitatively assessing the quality of the barcodes.
   Specifically, one cannot just say I'll disregard the "shorter" ones, and the left are the features.
   How short? What about the variation in length?

2. The space of barcodes lacks geometric properties to define basic concepts Mean, median, etc.

2.1.1 Persistence Diagram

Few tools can be used in applications for computation of Wasserstein / bottleneck distance between persistence diagrams.
Some existing ones include: Dionysus, Hera, TDA Package.

3 Questions

3.1 Theory

3.1.1 Homology

1. What other "features" can be revealed by homology?
   It seems that although Homology groups are informational
   Betti numbers, which are the basis for barcode and other diagrams, do not tell us anything more that holes and voids.

3.2 Computation

1. What is the bottleneck of the general computation process?
   (a) Is it the same for all the libraries?
   (b) Possible optimization? Meaningful?
2. What is the bottleneck for a specific data type?
3. For a specific purpose?

3.3 Application

1. Search for a problem?
   It seems that the methods is a not problem-solving-oriented.
   But rather developed from math intuition.
   Therefore seems bound to be less powerful than DL, which is clearly developed to solve problems.

4 TODO List

4.1 General Learning Plan

1. Need to read the roadmap[1] more details.

4.2 Classic Dimensionality Reduction

1. Learn about the following:
   (a) PCA
   (b) MDS
   (c) isomap
2. Maybe need to implement some of them
3. Know the difference from a practical aspect

4.3 Clustering

1. Need to learn about the different approaches
   (a) Variational
   (b) Spectral
   (c) Hierarchical
   (d) Density thresholding
   (e) Mode seeking
   (f) Valley Seeking
2. Need to know the specific applications
3. Need to delve into one of them to see if possible for improvement
4. Mathematical details about Mode Seeking
   (a) Hypotheses on the function of f and estimator $\hat{f}$
   (b) Exact conclusions about the prominence gap.
   (c) How to specify prominence parameter τ
   (d) Algorithm details

4.4 Construct Complex

1. Need to learn about the various complexes, esp. their advantages / disadvantages for applications.
   (a) Čech
   (b) VR
   (c) Delaunay
   (d) Alpha
   (e) Witness

4.5 Persistent Homology

1. How to compute a filtration from point cloud?
   (a) Various ways to define simplices, Cech, VR, etc
   (b) What if data not necessarily point cloud?
      i. Image?
      ii. 3D objects?
2. Need to delve into the details of the PH algorithms
   (a) Like reading off the intervals.
   (b) Different implementations.
3. ALL the above pretty much FINISHED.

5 Clustering

Point cloud (with coordinates)

Distance / dissimilarity matrix

Note: this seems to be a different idea of the distance function used in TDA, which is called lens in that context.

Barcode →merge tree →dendrogram

5.1 Mode Seeking Paradigm

Problems:

1. Noisy estimator
2. Neighborhood graph

Solutions:

1. Be proactive: smooth
2. Be reactive: merge clusters after clustering
   This leads to "topological persistence"

Persistence for Model Seeking:
Probability density function f

- Nested family (filtration) of of inverse images, or superlevel-sets $f^{-1}([t, +\infty))$ for t from $+\infty$ to $-\infty$
- Track evolution of "topology"

Similar stability theorem.
Seems to have relation with Morse theory. "If f is Morse, then..."

6 Topological Persistence

Persistence diagram shows the "persistence" of the topological features. Slight perturbation causes slight difference in persistence diagrams.

7 Homology

Definition: $h : X \mapsto Y$ is a homeomorphism if there exists a map $h^{-1} : Y \mapsto X$, s.t.

- both are continuous
-

8 How to deal with data

8.1 Networks

Essentially 1-dimensional simplicial complex. Filter by weight.
Construct higher dimensional simplices. Ex: WRCF.
Mapping the nodes to points in finite metric space.

8.2 Digital Images

Natural cubical structure. Cubical complexes.
c color variables, N pixels/voxels, then $c \times N$-dimensional space, equipped a distance function to form a finite metric space.

9 Construct (Simplicial) Complex

9.0.1 Various Ways

9.0.2 Reduction Techniques

Heuristic ways to reduce the size of a filtered complex while keeping the PH unchanged.

Discrete Morse Theory Refer to [5] for theoretic details. NP complete, thus relies on heuristics to find partial matching to reduce the size.

Strong Collapses Refer to [12] for details.

10 Computation of Persistence Homology

This part is largely cited from [1].
How to keep track of how one feature "merges" to another?
Boundary matrix: the matrix representation of boundary operator.
We also need a total ordering compatible with the "filtration" in the following sense:

- a face of a simplex precedes the simplex.
- a simplex in K_i precedes simplices in K_j for $j > i$, and not in K_i.
  - this essentially means that we place the simplices by the order of "appearing".

10.1 Standard Algorithm

- Form the boundary matrix from the ordering.
- Reduction, which is essentially Gaussian elimination.
- Reading off intervals.
  1. some details to do.
  2. degree: $\mathrm{dg}(\sigma) =$ smallest number l s.t.$\sigma \in K_l$
  3. pair (σ_i, σ_j) gives $[\mathrm{dg}(\sigma_i), (\sigma_j))$
  4. unpaired extends to infinity.
-

10.2 Complexity

In the worst case, which does exist, the algorithm has cubic complexity. Note that when sparse, not cubic.

11 Statistical Interpretation

11.1 Problems

1. Compare outputs with null model.
   How to compare the different results?
   How to evaluate the significance of the data?
2. Average over multiple realizations of a random model.

11.2 Statistical Analysis

Statistical methods for PH addressed first time in [6]. Three approaches.

1. Topological properties of random simplicial complexes, viewed as null models.
2. Properties of a metric space, whose points are persistence diagrams.
3. "Features" of persistence diagrams.

11.2.1 Second Approach

Key point: define an appropriate distance function between diagrams. The main object is the persistence diagram, which is in some sense isomorphic to a barcode.

Wasserstein Distance is a popular way to defined distance. Details skipped.

11.2.2 Third Approach

Key point: map the space of persistence diagrams to analyzable spaces (e.g. Banach spaces). Persistence landscape[13], using space of algebraic functions[14], kernalized techniques.

12 Mapper

13 Why TDA?

13.1 Theoretically well understood

13.2 Qualitative data features

13.3 Computable via linear algebra

13.4 Robust under perturbation

14 Application: Possible Topics and Data Sets to Try

14.1 List of fields

1. Sensor network coverage.
2. Proteins.
3. 3-dimensional structure of DNA
4. Robotics
5. Signals in images
6. Periodicity in time series
7. Cancer
8. Phylogenetics
9. Natural images
10. Self-similarity in geometry
11. Materials science
12. Financial networks
13. Neuroscience
14. Other networks
15. Time-series output of dynamical systems
16. Natural language analysis

14.2 Important Examples

14.2.1 Digital Image from Otter 17

This example comes from [1].
Two approaches: Cubical Complexes and //TODO

Cubical Complexes Instead of simplices, one use cubical complexes to build the topological space.
Sounds like Čech?

1. Every pixel is a point, join the adjacent points;
2. Color the unit squares and cubes (only unit ones!);
3. Label each pixel with grey scale, each edge with the max of point;
4. Filter the topology by grey scale values;

14.3 Data Sets

14.4 Choice of Software

List of some open source softwares available

1. Javaplex[9]
2. Dionysus
3. DIPHA
4. Perseus
5. GUDHI
6. Ripser, the known best one, according to [1]

References

[1] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," *EPJ Data Science*, vol. 6, p. 17, Aug 2017.

[2] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc. (N.S.)*, vol. 46, no. 2, pp. 255–308, 2009.

[3] G. Singh, F. Memoli, and G. Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition," in *Eurographics Symposium on Point-Based Graphics* (M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, eds.), The Eurographics Association, 2007.

[4] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, "On the local behavior of spaces of natural images," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 1–12, 2008.

[5] K. Mischaikow and V. Nanda, "Morse theory for filtrations and efficient computation of persistent homology," *Discrete & Computational Geometry*, vol. 50, pp. 330–353, Sep 2013.

[6] P. Bubenik and P. T. Kim, "A statistical approach to persistent homology," *Homology Homotopy Appl.*, vol. 9, no. 2, pp. 337–362, 2007.

[7] H. Wagner, C. Chen, and E. Vuçini, *Efficient Computation of Persistent Homology for Cubical Data*, pp. 91–106. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[8] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational homology*, vol. 157 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[9] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A research software package for persistent (co)homology," in *Proceedings of ICMS 2014* (H. Hong and C. Yap, eds.), Lecture Notes in Computer Science 8592, pp. 129–136, 2014. Software available at `http://appliedtopology.github.io/javaplex/`.

[10] B. T. Fasy, J. Kim, F. Lecci, and C. Maria, "Introduction to the R package TDA," *CoRR*, vol. abs/1411.1830, 2014.

[11] P. Bubenik and P. Dłotko, "A persistence landscapes toolbox for topological statistics," *Journal of Symbolic Computation*, vol. 78, pp. 91 – 114, 2017. Algorithms and Software for Computational Topology.

[12] J. A. Barmak and E. G. Minian, "Strong homotopy types, nerves and collapses," *Discrete & Computational Geometry*, vol. 47, pp. 301–328, Mar 2012.

[13] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *J. Mach. Learn. Res.*, vol. 16, pp. 77–102, Jan. 2015.

[14] A. Adcock, E. Carlsson, and G. Carlsson, "The Ring of Algebraic Functions on Persistence Bar Codes," *ArXiv e-prints*, Apr. 2013.