Background Reading for Topological Data Analysis

Topological data analysis (TDA) is a methodology for applying the mathematical study of shape (topology) to the study of large and complex data sets. Its development was initiated around the year 2000. It has now seen applications in a large number of varied domains. The purpose of this document is to give a brief summary of the literature and outline of the directions in which the methodology has been applied.

There are two distinct threads in the subject. One is the *mapping* thread, which produces shapes encoded as graphs or simplicial complexes on which one can operate directly. The second thread is the *shape measurement* thread, in which one can construct invariants of shapes that can be used to obtain understanding of data as well as to coordinatize various kinds of unstructured data. Both themes appear prominently in applications.

An overall survey of the area is given in [1]: G. Carlsson, *Topology and Data*, Bull. Am. Math. Soc. 46 (2009) 255-308

Mapping

This area is concentrated around one construction, given in [1] and [11], which is a generalization of the notion of Reeb graphs in computational geometry. It takes as input a point cloud (a data set equipped with a dissimilarity measure), and produces as output a network or graph in the computer science sense, which we refer to as the *topological model*. These graphs can be taken as maps describing the "similarity landscape" of the data. This kind of topological summary has been extremely useful in various applications. There is a theory developing around the stability of the construction, which will simplify and strengthen the ability to perform inference with it. Examples of this kind of work are [31] and [32]. In addition, the topological model as constructed by Ayasdi includes a great deal of functionality beyond the simple display of the map, which permits the development of models and applications deriving from the topological analysis, as well as inference of various kinds. Explicit comparisons with some other methods for unsupervised analysis have been performed in [21] in the context of hyperspectral imaging. The published research has been concentrated in the biomedical realm, although the range of applications is growing beyond that. The applications include work in numerous areas, enumerated below with the corresponding references.

Cancer genomics: [4], [12], [27]
Genetics: [12], [14], [16]
Infectious Disease: [2], [3], [15]
Asthma: [19], [20]
Diabetes: [41]
Autism related syndrome: [18]
Chemical Imaging: [21]
Traumatic Brain Injury: [7], [8]

Shape Measurement

The work here is concentrated around *persistent homology*. This method is an adaptation of the homology signatures in standard algebraic topology, that are able to capture the presence of various kinds of patterns in shapes. A survey of the method is given in [10]. The output of this method is a signature called a barcode (or equivalently, a persistence diagram), which is somewhat analogous to the dendrograms produced by hierarchical clustering, but which capture higher order properties of a shape. There are two directions of applications of these signatures. One is the measurement of the overall shape of given data sets. This kind of application is carried out in [13] (viral evolution), [22] and [42] (image processing), and [35] and [36] (neuroscience). Applications of this kind of analysis includes the construction of compression schemes [26] and coordinatization of texture data [28]. The second direction is the coordinatization of unstructured data, such as databases of molecules or images. This kind of application is exemplified by the work described in [33] and [43], and has been shown to be quite effective in drug discovery and in the materials science of glasses and other materials. The technical background for both methods are given in [1], [10], [29], [30], and [40]. There is considerable theoretical work concerning stability properties of persistent homology, for example in [23] and [24]. The work describing coordinatization methods based on persistent homology is exemplified in [38] and [39].

1. G. Carlsson, *Topology and data*, Bull. Am. Math. Soc. 46 (2009) 255-308
2. B.Y. Torres, J.H.M. Oliveira, A.T. Tate, P. Rath, K. Cumnock, D.S. Schneider, Tracking resilience to infections by mapping disease space, PLoS Biol. 14 (2016) e1002436, https://doi.org/10.1371/journal.pbio.1002436.
3. A. Louie, K.H. Song, A. Hotson, A. Thomas Tate, D.S. Schneider, How many parameters does it take to describe disease tolerance? PLoS Biol. 14 (2016) e1002435, https://doi.org/10.1371/journal.pbio.1002435.

4. J.-K. Lee, J. Wang, J.K. Sa, E. Ladewig, H.-O. Lee, I.-H. Lee, H.J. Kang, D.S. Rosenbloom, P.G. Camara, Z. Liu, P. van Nieuwenhuizen, S.W. Jung, S.W. Choi, J. Kim, A. Chen, K.-T. Kim, S. Shin, Y.J. Seo, J.-M. Oh, Y.J. Shin, C.- K. Park, D.-S. Kong, H.J. Seol, A. Blumberg, J.-I. Lee, A. Iavarone, W.-Y. Park, R. Rabadan, D.-H. Nam, Spatiotemporal genomic architecture informs precision oncology in glioblastoma, Nat. Genet. 49 (2017) 594e599, https://doi.org/ 10.1038/ng.3806.
5. V. Pedoia, J. Haefeli, K. Morioka, H.-L. Teng, L. Nardo, R.B. Souza, A.R. Ferguson, S. Majumdar, MRI and biomechanics multidimensional data analysis reveals R 2 -R 1r as an early predictor of cartilage lesion progression in knee osteoarthritis: multidimensional Data Analysis to Study OA, J. Magn. Reson. Imaging (2017), https://doi.org/10.1002/jmri.25750.
6. G. Singh, F. Memoli, T. Ishkhanov, G. Sapiro, G. Carlsson, D.L. Ringach, Topological analysis of population activity in visual cortex, J. Vis. 8 (2008), https:// doi.org/10.1167/8.8.11, 11e11.
7. J.L. Nielson, J. Paquette, A.W. Liu, C.F. Guandique, C.A. Tovar, T. Inoue, K.- A. Irvine, J.C. Gensel, J. Kloke, T.C. Petrossian, P.Y. Lum, G.E. Carlsson, G.T. Manley, W. Young, M.S. Beattie, J.C. Bresnahan, A.R. Ferguson, Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury, Nat. Commun. 6 (2015) 8581, https://doi.org/10.1038/ ncomms9581
8. J.L. Nielson, S.R. Cooper, J.K. Yue, M.D. Sorani, T. Inoue, E.L. Yuh, P. Mukherjee, T.C. Petrossian, J. Paquette, P.Y. Lum, G.E. Carlsson, M.J. Vassar, H.F. Lingsma, W.A. Gordon, A.B. Valadka, D.O. Okonkwo, G.T. Manley, A.R. Ferguson, TRACKTBI Investigators, Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis, PLoS One 12 (2017) e0169490, https://doi.org/10.1371/journal.pone.0169490.
9. P.Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, G. Carlsson, Extracting insights from the shape of complex data using topology, Sci. Rep. 3 (2013), https://doi.org/10.1038/ srep01236.
10. G. Carlsson, Topological pattern recognition for point cloud data, Acta Numer. 23 (2014) 289e368, https://doi.org/10.1017/S0962492914000051.
11. G. Singh, F. Memoli, G.E. Carlsson, Topological methods for the analysis of high-dimensional data sets and 3d object recognition, SPBG (2007) 91e100
12. P.G. Camara, Topological methods for genomics: present and future directions, Curr. Opin. Syst. Biol. 1 (2017) 95e101, https://doi.org/10.1016/ j.coisb.2016.12.007.
13. J.M. Chan, G. Carlsson, R. Rabadan, Topology of viral evolution, Proc. Natl. Acad. Sci. 110 (2013) 18566e18571.
14. K.J. Emmett, R. Rabadan, Characterizing scales of genetic recombination and antibiotic resistance in pathogenic bacteria using topological data analysis, in: Int. Conf. Brain Inform. Health, Springer, 2014, pp. 540e551.
15. A.M. Ibekwe, J. Ma, D.E. Crowley, C.-H. Yang, A.M. Johnson, T.C. Petrossian, P.Y. Lum, Topological data analysis of Escherichia coli O157:H7 and non-O157 survival in soils, Front. Cell. Infect. Microbiol. 4 (2014), https://doi.org/ 10.3389/fcimb.2014.00122.
16. P.G. Camara, A.J. Levine, R. Rabadan, Inference of ancestral recombination graphs through topological data analysis, PLoS Comput. Biol. 12 (2016) e1005071, https://doi.org/10.1371/journal.pcbi.1005071.

17. P.G. Camara, D.I.S. Rosenbloom, K.J. Emmett, A.J. Levine, R. Rabadan, Topological data analysis generates high-resolution, genome-wide maps of human recombination, Cell Syst. 3 (2016) 83e94, https://doi.org/10.1016/ j.cels.2016.05.008.
18. D. Romano, M. Nicolau, E.-M. Quintin, P.K. Mazaika, A.A. Lightbody, H. Cody Hazlett, J. Piven, G. Carlsson, A.L. Reiss, Topological methods reveal high and low functioning neuro-phenotypes within fragile X syndrome, Hum. Brain Mapp. 35 (2014) 4904e4915, https://doi.org/10.1002/hbm.22521.
19. T.S.C. Hinks, X. Zhou, K.J. Staples, B.D. Dimitrov, A. Manta, T. Petrossian, P.Y. Lum, C.G. Smith, J.A. Ward, P.H. Howarth, A.F. Walls, S.D. Gadola, R. Djukanovic, Innate and adaptive T cells in asthmatic patients: relationship to severity and disease mechanisms, J. Allergy Clin. Immunol. 136 (2015) 323e333, https://doi.org/10.1016/j.jaci.2015.01.014.
20. T.S.C. Hinks, T. Brown, L.C.K. Lau, H. Rupani, C. Barber, S. Elliott, J.A. Ward, J. Ono, S. Ohta, K. Izuhara, R. Djukanovic, R.J. Kurukulaaratchy, A. Chauhan, P.H. Howarth, Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and 130 L. Duponchel / Analytica Chimica Acta 1000 (2018) 123e131 chitinase 3elike protein 1, J. Allergy Clin. Immunol. 138 (2016) 61e75, https:// doi.org/10.1016/j.jaci.2015.11.020.
21. M. Offroy, L. Duponchel, Topological data analysis: a promising big data exploration tool in biology, analytical chemistry and physical chemistry, Anal. Chim. Acta 910 (2016) 1e11, https://doi.org/10.1016/j.aca.2015.12.037.
22. G. Carlsson, T. Ishkhanov, V. de Silva and A. Zomorodian (2008), ‘On the local behavior of spaces of natural images’, Internat. J. Computer Vision 76, 1–12.
23. F. Chazal, D. Cohen-Steiner, L. Guibas, F. Memoli and S. Oudot (2009), Gromov–Hausdorff stable signatures for shapes using persistence. In Eurographics Symposium on Geometry Processing 2009. Computer Graphics Forum 28, 1393– 1403.
24. D. Cohen-Steiner, H. Edelsbrunner and J. Harer (2007), ‘Stability of persistence diagrams’, Discrete Comput. Geom. 37, 103–120.
25. H. Edelsbrunner, D. Letscher and A. Zomorodian (2002), ‘Topological persistence and simplification’, Discrete Comput. Geom. 28, 511–533
26. A. Maleki, M. Shahram and G. Carlsson (2008), Near optimal coder for image geometries. In Proc. 15th IEEE International Conference on Image Processing (ICIP), pp. 1061–1064.
27. M. Nicolau, A. Levine and G. Carlsson (2011), Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc. Nat. Acad. Sci. 108, 7265–7270.
28. J. Perea and G. Carlsson (2014), ‘A Klein bottle-based dictionary for texture representation’, Internat. J. Computer Vision 107, 75–97.
29. A. Zomorodian and G. Carlsson (2005), ‘Computing persistent homology’, Discrete Comput. Geom. 33, 247–274.
30. A. Zomorodian (2005), Topology for Computing, Vol. 16 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press.
31. Carrière, M. & Oudot, S. *Structure and stability of the one-dimensional mapper*, Found Comput Math (2017). https://doi.org/10.1007/s10208-017-9370-z
32. de Silva, V., Munch, E. & Patel, A. *Categorified Reeb graphs*, Discrete Comput Geom (2016) 55: 854. https://doi.org/10.1007/s00454-016-9763-9

33. Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G. Escolar, Kaname Matsueand Yasumasa Nishiura, *Hierarchical structures of amorphous solids characterized by persistent homology*, Proc. Natl. Acad. Sciences, June 2016, 113 (26) 7035-7040. https://doi.org/10.1073/pnas.1520877113
34. Y. Yao, J. Sun, X. Huang, G. Bowman, G. Singh, M. Lesnick, L. Guibas, V. Pande, G. Carlsson, *Topological methods for exploring low-density states in biomolecular folding pathways*, J. Chem. Physics, 2009, 130 (14), 144115. doi: 10.1063/1.3103496.
35. Bendich, P., Marron, J., Miller, E., Pieloch, A. and Skwerer, S. (2014). Persistent homology analysis of brain artery trees. Ann. Appl. Stat. 10(1): 198-218.
36. Giusti, C., Ghrist, R. and Bassett, D. (2016) Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. J. Comput. Neurosci. 41 (1): doi:10.1007/s10827-016-0608-6
37. E. Lazar, J. Mason, R. MacPherson, and D. Srolovitz, *Complete topology of cells, grains, and bubbles in three-dimensional microstructures*, Phys. Rev. Letters, 2012,109(9).
38. P. Bubenik, *Statistical topological data analysis using persistence landscapes*, Journal of Machine Learning Research 16 (2015) 77-102.
39. A. Adcock, E. Carlsson, and G. Carlsson, *The ring of algebraic functions on persistence bar codes,* Homology, Homotopy, and Applications, 18(1), 2016, 381-402
40. H. Edelsbrunner and J. Harer, **Computational Topology: and Introduction**, American Mathematical Society, 2010.
41. Li Li, W. Cheng, B. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. Bottinger, and J. Dudley, *Identification of type 2 diabetes subgroups through topological analysis of patient similarity,* Science Translational Medicine Vol. 7, Issue 311, 2015, DOI: 10.1126/scitranslmed.aaa9364
42. H. Adams and G. Carlsson, *On the nonlinear statistics of range image patches,* SIAM Journal of Imaging Sciences, Vol. 2, no. 1, 110-117.
43. G. Wei, *Persistent homology analysis of biomolecular data,* SIAM news, December 2017.