Preface

Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik and Marius Thaule

The demands of science and industry for methods for understanding and utilizing large and complex data sets have been growing very rapidly, driven in part by our ability to collect ever more data about many different subjects. A key requirement is to construct useful models of data sets that allow us to see more clearly and rapidly what the data tells us. Mathematical modeling is usually thought of as the discipline of constructing *algebraic* or *analytic* models, where the output of the model is an equation, a system of equations, or perhaps a system of differential equations. This method has been very effective in the past, when many of the data sets to be studied involved only a small number of features and where there are simple relations among the variables that govern the data being modeled. The work of Galileo, Kepler, and Newton are prime examples of the successes of this kind of modeling. However, these methods run into difficulties when confronted with some of the very complex data currently arising in applications. For example, consider data sets where the goal is to identify potential instances of fraud, or to discover drugs, where the

Nils A. Baas
Deptartment of Mathematical Sciences, NTNU, NO-7491 Trondheim, Norway, e-mail: nils.baas@ntnu.no

Gunnar E. Carlsson
Department of Mathematics, Stanford University, Stanford, California 94305, USA, e-mail: carlsson@stanford.edu

Gereon Quick
Deptartment of Mathematical Sciences, NTNU, NO-7491 Trondheim, Norway, e-mail: gereon.quick@ntnu.no

Markus Szymik
Deptartment of Mathematical Sciences, NTNU, NO-7491 Trondheim, Norway, e-mail: markus.szymik@ntnu.no

Marius Thaule
Deptartment of Mathematical Sciences, NTNU, NO-7491 Trondheim, Norway, e-mail: marius.thaule@ntnu.no

complex structure of molecules means that identification of effective medications is a very complex task. For this reason, it is incumbent on the mathematical and statistical communities to develop new methods of modeling. To understand what these methods might be, we ask ourselves what do mathematical models buy us? Here are some answers to that question.

- A mathematical model should provide some kind of compression of the data into a tractable form. When we model data by using a simple one variable linear regression, the result compresses the data from thousand or hundreds of thousands of data points into two numbers, the slope and the y-intercept. If the approximation is good, we have achieved a massive compression.
- A mathematical model should provide understanding of the data. The usual mathematical modeling of the flight of a cannonball gives a great deal of understanding about its behavior.
- In many cases, we would like a model to allow us to predict outcomes. In the cannonball problem, we need only know the muzzle velocity and the angle of the cannon barrel in order to predict where the cannonball will land, or what the highest altitude it will reach is.

Nothing about these answers requires that the model be algebraic. Consider, for example, cluster analysis. Its output is no longer an equation or a set of equations, but rather a partition of the data set into a collection of groups. Such a partition provides all three of the capabilities described above. Cluster analysis clearly provides compression, since the number of clusters is typically a much smaller number than the number of data points. It also provides understanding, since the cluster decomposition is effectively a taxonomy of the data points. Finally, it can also be used to provide predictions, via classifying new data points into the different clusters using methods like logistic regression or decision trees. These observations suggest that we view cluster analysis as a modeling mechanism which is discrete in the sense that it produces zero-dimensional outputs, with no information about continuous phenomena such as progressions. They also suggest that we look for other modeling mechanisms where the output can consist of more complex mathematical structures. Topological data analysis (TDA) is a modeling method in which the outputs are graphs and simplicial complexes. Work on TDA began with the study of *persistent homology* (see [16, 26, 32]), but over time the direct study of low-dimensional simplicial complex models (see [4, 30]) has also become important in applications. Here are some of the advantages of TDA.

- TDA is able to give insight into continuous *and* discrete properties of a data set in one output. Cluster analysis provides a discrete analysis, and algebraic modeling often reflects continuous information.

- It is able to represent the properties of complex data more flexibly and therefore more accurately than other machine learning methods.
- There is a great deal of “functionality" in the representation of data sets, since simplicial complexes and graphs are more complex mathematical structures than partitions or simple regression models. For example, if one is studying a function on a data set, one is often able to create a corresponding function on the nodes of the model, and the behavior of the corresponding function often clarifies the behavior of the function. Persistent homology can also be viewed as functionality, since it provides a way to measure (in an appropriate sense) the shape of the model.
- An interesting direction is the study of topological models of the set of features in a data set rather than the set of data points. This point of view has been advocated in [27] and [11], and referred to in [27] as “topological signal processing".
- Although persistent homology can be used to study the overall structure of data sets, it is also used to generate features of data sets of complex or unstructured objects. For example, in [31], data bases of molecules are treated as data sets whose points are finite metric spaces.

TDA has been applied in a number of interesting domains, notably neuroscience [18, 20, 25, 29, 28], materials science [19, 22], cancer biology [21, 23], and immune responses [24].

There are numerous very active mathematical research directions within TDA.

- **Vectorization of barcodes:** Most machine learning methods are defined for data which is in the form of vectors in a high dimensional vector space. There are numerous situations where the data points themselves are more complex objects, which support a metric. For example, molecule structures or images fall into this category. In such situations, one has assignments of barcodes to individual data points instead of the whole data set. In order to enable machine learning, one must therefore create functions on the set of barcodes. There are a number of strategies to provide such “vectorizations". See [1, 2, 8] for examples.
- **Probabilistic analysis of spaces of barcodes:** Statistical and probabilistic analyses clearly play a key role in any data analytic problem. If we are building simplicial complex models or creating features based on persistent homology, it is clear that it is important to understand the behavior of distributions on the set (it can be made into a metric space in numerous ways) of persistence barcodes or equivalently persistence diagrams. There is a great deal of work in this direction. See [3, 5, 6, 7, 15] for interesting examples.
- **Methods for assessing the faithfulness of topological models:** If we build topological models of data, it is critical to devise methods for assessing how faithful to the data the model is. Of course, even the problem of defining measures

of this kind of consistency is an important one. The paper [12] is an example of this kind of work.

- **Multidimensional and generalized persistence:** Since the development of persistent homology, a number of generalizations of it have been developed. In particular, the idea that one might have families of complexes depending on more than one real parameter is referred to as *multidimensional persistence* [9]. Additionally, *zig-zag persistence* [10] studies the behavior of parametrized families of complexes where one is permitted to delete as well as add simplices. Further generalizations have been made, and a key direction of research is to attach invariants to generalized persistence objects so that one can interpret them and make use of them in data analysis. Other interesting work in this direction is given in [13, 17].
- **New domains of application:** TDA has already seen application in numerous areas, which were mentioned above. Finding new ways to apply it is high priority research.

This volume presents a number of interesting papers in numerous different research directions. It provides a partial snapshot of the current state of the field, and we hope that it will be useful to practitioners as well as those considering entering the field.

The papers are written by participants (and their collaborators) of the Abel Symposium 2018 which took place from June 4 to June 8, 2018 in Geiranger, Norway. The symposium was organized by an external committee consisting of Gunnar E. Carlsson (Stanford University), Herbert Edelsbrunner (IST Austria), Kathryn Hess (EPF Lausanne), and Raul Rabadan (Columbia University) and a local committee from NTNU Trondheim consisting of Nils A. Baas, Gereon Quick, Markus Szymik and Marius Thaule. The webpage of the symposium can be found at https://folk.ntnu.no/mariusth/Abel/.

We gratefully acknowledge the generous support of the Board for the Niels Henrik Abel Memorial Fund, the Norwegian Mathematical Society, the Department of Mathematical Sciences and the Faculty of Information Technology and Electrical Engineering at NTNU. We also thank Ruth Allewelt, Leonie Kunz and Springer-Verlag for encouragement and support during the editing of these proceedings.

Stanford and Trondheim
October 2019

Nils A. Baas
Gunnar E. Carlsson
Gereon Quick
Markus Szymik
Marius Thaule

References

1. Adams H., Emerson T., Kirby M., Neville R., Peterson C., Shipman P., Chepushtanova S., Hanson E., Motta F., Ziegelmeier L.: *Persistence images: a stable vector representation of persistent homology*. J. Machine Learning Research **18**, 1–35 (2017)
2. Adcock A., Carlsson E., Carlsson G.: *The ring of algebraic functions on persistence barcodes*. Homology, Homotopy, and Applications **18**, 381–402 (2016)
3. Adler R., Taylor J.: Random Fields and Geometry. Springer (2009)
4. Akkiraju N., Edelsbrunner H., Facello M., Fu P., Mucke E., Varela C.: *Alpha shapes: defintion and software*. In: Proc. Internat. Comput. Geom. Software Workshop 1995
5. Blumberg A., Gal I., Mandell M., Pancia M.: *Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces*. Foundations of Computational Mathematics **14**, 745–789 (2014)
6. Bobrowski O., Borman M.: Euler integration of Gaussian random fields and persistent homology. Journal of Topology and Analysis **4**, 49–70 (2012)
7. Bobrowski O., Kahle M., Skraba P.: *Maximally persistent cycles in random geometric complexes*. Annals of Applied Probability **27**, 2032–2060 (2017)
8. Bubenik P.: *Statistical topological data analysis using persistence landscapes*. The Journal of Machine Learning Research **16**, 77–102 (2015)
9. Carlsson G., Zomorodian A.: The theory of multidimensional persistence. Discrete and Computational Geometry **42**, 71–93 (2009)
10. Carlsson G., V. de Silva V.: Zigzag persistence. Foundations of Computational Mathematics **10**, 367–405 (2010)
11. Carlsson G., Gabrielsson R.B.: Topological approaches to deep learning. These proceedings 2019
12. Carrière M., S. Oudot S.: Structure and stability of the one-dimensional Mapper. Foundations of Computational Mathematics **18**, 1333–1396 (2018)
13. Chacholski W., Scolamiero M., VaccarinoF.: Combinatorial presentation of multidimensional persistent homology. J. Pure and Applied Algebra **221**, 1055–1075 (2017)
14. Chan J., Carlsson G., Rabadan R.: Topology of viral evolution. PNAS **110** (46), 18566–18571 (2013) https://doi.org/10.1073/pnas.1313480110
15. Chazal F., Fasy B., Lecci F., Michel B., Rinaldo A., Wasserman L.: Robust topological inference: distance to a measure and kernel distance. Journal of Machine Learning Research **18**, 1–40 (2018)
16. Edelsbrunner H., Letscher D., Zomorodian A.: Topological persistence and simplification. Discrete and Computational Geometry **28**, 511–533 (2002)
17. Cagliari F., B. Di Fabio B., Ferri M.: One-dimensional reduction of multidimensional persistent homology. Proc. Amer. Mat. Soc. **138**, 3003–3017 (2010)
18. Giusti C., Pastalkova E., Curto C., V. Itskov V.: Clique topology reveals intrinsic geometric structure in neural correlations. PNAS **112** (44), 13455–13460 (2015) https://doi.org/10.1073/pnas.1506407112
19. Hiraoka Y., Nakamura T., Hirata A., Excolar E.G., Matsue K., Nishiura Y.: Hierarchical structures of amorphous solids characterized by persistent homology. PNAS **113** (26), 7035–7040 (2016) https://doi.org/10.1073/pnas.1520877113
20. Kanari L., Dlotko P., Scolamiero M., Levi R., Shillcock J.C., Hess K., Markram H.: A topological representation of branching morphologies. Neuroinformatics (2017)
21. Lee J-K., et al: Spatiotemporal genomic architecture informs precision oncology in glioblastoma. Nature Genetics **49**, 594–599 (2017)
22. MacPherson R., Schweinhart B.: Measuring shape with topology. J. Math. Phys. **53** (2012)
23. Nicolau M., Levine A., Carlsson G.:Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. PNAS **108** (17), 7265–7270 (2011) doi: 10.1073/pnas.1102826108

24. Olin A., Henckel E., Chen Y., Lakshmikanth T., Pou C., Mikes J., Gustafsson A., Bernhardsson A., Zhang C., Bohlin K., Brodin P.: Stereotypic immune system development in newborn children. Cell, 2018 Aug 23; **174** (5), 1277–1292 (2018) doi: 10.1016/j.cell.2018.06.045
25. Reimann M.W., Nolte M., Scolamiero M., Turner K., Perin R., Chindemi G., Dlotko P., Levi R., Hess K., Markram H.: Cliques of neurons bound into cavities provide a missing link between structure and function. Front. Comput. Neurosci. (2017)
26. Robins V.: Towards computing homology from finite approximations. Proceedings of the 14th Summer Conference on General Topology and its Applications (Brookville, NY, 1999), Topology Proc. 24, 1999, 503–532 (1999)
27. Robinson M.: Topological Signal Processing. Springer Verlag (2014)
28. Rybakken E., Baas N., Dunn B.: Decoding of neural data using cohomological features extraction. Neural Computation **31**, 68–93 (2019)
29. Saggar M., Sporns O., Gonzalez-Castillo J., Bandettini P., Carlsson G., Glover G., Reiss R.: Towards a new approach to reveal dynamical organization of the brain using topological data analysis. Nature Communications **9** Article number 1399 (2018)
30. Singh G., Memoli F., Carlsson G.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: Eurographics Symposium on Point-Based Graphics (2007)
31. Xia K., Wei G.: Persistent homology analysis of protein structure, flexibility and folding. International Journal for Numerical Methods in Biomedical Engineering **30**, 814–844 (2014)
32. Zomorodian A., Carlsson G.: Computing persistent homology, Discrete and Computational Geometry. **33**, 249–274 (2005)