Topological Data Analysis with Applications

The continued and dramatic rise in the size of data sets has meant that new methods are required to model and analyze them. This timely account introduces topological data analysis (TDA), a method for modeling data by geometric objects, namely graphs and their higher-dimensional versions, simplicial complexes. The authors outline the necessary background material on topology and data philosophy for newcomers, while more complex concepts are highlighted for advanced learners. The book covers all the main TDA techniques, including persistent homology, cohomology, and Mapper. The final section focuses on the diverse applications of TDA, examining a number of case studies ranging from monitoring the progression of infectious diseases to the study of motion capture data.

Mathematicians moving into data science, as well as data scientists or computer scientists seeking to understand this new area, will appreciate this self-contained resource which explains the underlying technology and how it can be used.

Gunnar Carlsson is Professor Emeritus at Stanford University. He received his doctoral degree from Stanford in 1976, and has taught at the University of Chicago, at the University of California, San Diego, at Princeton University, and, since 1991, at Stanford University. His work within mathematics has been concentrated in algebraic topology, and he has spent the last 20 years on the development of topological data analysis. He is also passionate about the transfer of scientific findings to real-world applications, leading him to found the topological data analysis-based company Ayasdi in 2008.

Mikael Vejdemo-Johansson is Assistant Professor in the Department of Mathematics at City University of New York, College of Staten Island. He received his doctoral degree from Friedrich-Schiller-Universität Jena in 2008, and has worked in topological data analysis since his first postdoc with Gunnar Carlsson at Stanford 2008–2011. He is the chair of the steering committee for the Algebraic Topology: Methods, Computation, and Science (ATMCS) conference series and runs the community web resource appliedtopology.org.

Topological Data Analysis with Applications

GUNNAR CARLSSON
Stanford University, California

MIKAEL VEJDEMO-JOHANSSON
City University of New York, College of Staten Island and the Graduate Center

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University’s mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108838658
DOI: 10.1017/9781108975704

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Carlsson, G. (Gunnar), 1952– author. | Vejdemo-Johansson, Mikael, 1980– author.
Title: Topological data analysis with applications / Gunnar Carlsson, Mikael Vejdemo-Johansson.
Description: New York : Cambridge University Press, 2021. |
Includes bibliographical references and index.
Identifiers: LCCN 2021024970 | ISBN 9781108838658 (hardback)
Subjects: LCSH: Topology. | Mathematical analysis. | BISAC: MATHEMATICS / Topology
Classification: LCC QA611 .C29 2021 | DDC 514/.23–dc23
LC record available at https://lccn.loc.gov/2021024970

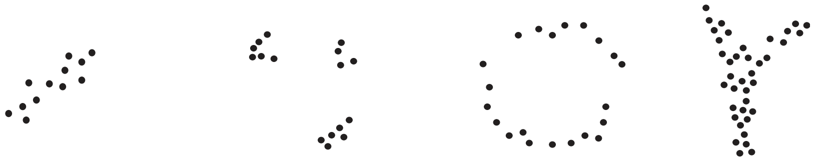
ISBN 978-1-108-83865-8 Hardback

Cambridge University Press
978-1-108-83865-8 — Topological Data Analysis with Applications
Gunnar Carlsson , Mikael Vejdemo-Johansson
Frontmatter
More Information

Contents

Preface

Data sets come in many shapes and sizes. The data sets presented in the figure below illustrate this point very well.



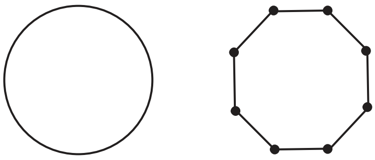
- The data set on the far left has the rough shape of a line in the plane. We are all familiar with many examples of this kind of data, and such data are typically handled with various regression models, which permit prediction and also allow for greater understanding of the data. This helps in developing mental models.
- The second set from the left illustrates a data set which decomposes into disjoint groups, and is not well approximated by any line. This kind of data occurs very frequently in the biomedical and social sciences, and cluster analysis has been developed to produce such decompositions in order to deliver taxonomies for the data.
- The third set from the left is a type of data set that occurs frequently when one is dealing with time series data representing periodic or recurrent behavior of some kind.
- The data set on the far right might describe data in which there are one standard or normal mode and three extremal modes. For example, it might come from sensors on an airliner, where the standard mode is flying at altitude in non-turbulent conditions and where the three extremal modes are takeoff, landing, and flying at altitude in turbulent conditions.

The first two data sets have dedicated methodologies (regression and cluster analysis, respectively) for their analysis. The latter two do not, although we believe that one could develop such methods for each of these two types. However, since these four data sets are by no means a complete list of the possible shapes of data, we can rapidly convince ourselves that hoping to create a complete list of shapes with tailor-made methods for each data shape is not the best solution to the problem dealing with all the different complexities that we can expect to find.

A possible approach to modeling data sets like those above is to view a modeling mechanism as a way to approximate data by sets with various shapes. For example,

Cambridge University Press
978-1-108-83865-8 — Topological Data Analysis with Applications
Gunnar Carlsson , Mikael Vejdemo-Johansson
Frontmatter
More Information

linear regression is the approximation of data by lines, planes, etc., while cluster analysis can be viewed as the approximation of data by finite discrete sets of points. Using this approach, the third data set from the left in the figure above could be approximated by a loop, and the data set on the far right could be thought of as having the shape of a letter “Y”, i.e., of three line segments which all join at a single central point. What these observations suggest is that we should develop a single method that can represent *all* shapes in one package. Fortunately, the mathematical discipline called topology provides exactly such a method. It turns out that graphs (and somewhat more complex objects called simplicial complexes) are very useful ways to describe shapes.



In this figure, we see a circle on the left and an octagon on the right. The two are very similar, in the sense that the octagon approximates the circle well, although it does not capture the curvature. There is a highly non-linear parametrization in which the points of the circle encode points on the octagon, and vice versa, which is called a *homeomorphism*, and we regard the two as representing the same information.

However, the octagon can also be described by a purely combinatorial object, namely a list of its vertices and of its edges, together with the information about which vertex belongs to which edge. This kind of parametrization is called a triangulation of the circle, and is the key concept from topology for the study of data. In particular, we will develop methods for approximating data sets by shapes described in a combinatorial way in the same way as we described the octagon above. The relevant combinatorial objects are graphs (in the computer science or combinatorics sense) or objects called simplicial complexes, which include not only edges but higher-order subsets such as triangles, tetrahedra, and higher-dimensional analogues.

With the above discussion in mind, topological data analysis (TDA) can be summarized as the idea that data, like topological spaces, can be usefully modeled by combinatorial objects such as graphs and simplicial complexes. The subject has been developing rapidly over the last 20 years, and this volume is an attempt to describe the theory as well as a varied array of applications. The rationale for the development of these methods consists of a number of different observations concerning data science.

- Linear algebraic methods, such as principal component analysis or multidimensional scaling, because of their algebraic nature are often not flexible enough to capture complicated non-linearities in data and their scatterplot output is often not as informative as one would like. Simplicial complex models are more flexible and capable of expressing complexity in the data. They also admit a great deal of functionality, allowing for effective interrogation and search of the data, as detailed in our Section 4.3.5.

- Cluster analysis seeks to divide data into disjoint groups to create a taxonomy of the data set. In many situations where cluster analysis is applied, however, one finds that the natural output is not a partition of the data into disjoint groups but, rather, a “soft clustering” in which the data is broken into groups that may overlap. This kind of information is very naturally modeled with a simplicial complex, which is able to describe the relationships between groups implied by the overlaps, using an appropriate shape or space. An ordinary cluster decomposition, by a partition, is in this situation modeled by a zero-dimensional simplicial complex, i.e. a finite set of points.
- Data science is often confronted with the problem of deciding the appropriate shape for a data set, so as to be able to model it effectively. The theory of simplicial complexes is equipped with a method for describing the shape structure of the output of a model; this is based on an extension of the *homology* construction from the algebraic topology of spaces. The extension is called *persistent homology* and will be the subject of many of the ideas that we present.
- Feature selection and feature engineering is a major task in data science. It is particularly challenging for data sets which are unstructured in the sense that they are not well represented by a data matrix or a spreadsheet with numerical entries. For example, a database of large molecules is regarded as unstructured because it consists of an unordered set of atoms and an unordered set of bonds and, further, because the spatial coordinates of the atoms can be varied via a rigid motion of space while the structure of the molecule remains unchanged. This means that a representation by the coordinates of the atoms is not meaningful. It turns out, though, that the molecules themselves have a geometry expressed through interatomic distances, which allows us to apply the homology tools described above to generate meaningful numerical quantities that can be used for analysis. Images form another class of data which can be viewed as unstructured and which can be studied with homological methods.
- Another way in which topological methods can be used for feature engineering is the notion of topological signal processing (Robinson 2014). The idea here is that, when given a data matrix, it is also useful to develop a topological model for the columns of the data matrix (i.e. the features) rather than for the points or samples of the data (i.e. the rows). In this way, each of the original data points can be viewed as a function on the set of features and ultimately as a function on the topological model. Incorporating various methods, including graph Laplacians, one can impose structure on data points using this approach, and obtain topologically informed dimensionality reductions.

We believe that the use of TDA in data science will motivate interesting and useful developments within topology. It is therefore useful to see where TDA methods fit within standard algebraic topology and homotopy theory. Here are some important points about this fit.

- Persistent homology can be described as the study of diagrams whose shape is defined by the partially ordered set $\mathbb{R}$. A number of other diagrams have been studied, including those used in zig-zag persistence and multidimensional persistence. As the work in TDA broadens and deepens, it is likely that increasingly sophisticated diagrams will be useful for extracting more detailed information from data sets. It follows that the construction of invariants for diagrams of various shapes will be a useful endeavor.

- Because TDA operates by studying samples of discrete sets of points, the dimensions of spaces that can be analyzed solely by TDA methods are fairly low, for the most part ≤ 5. A data set which would faithfully represent a space of dimension 10 would be expected to require at least 10^{10} points, if one assumes a resolution of 10 points for each dimension. This is already a very large number, and demonstrates the point that, for example, 50-dimensional homology is not likely to occur in a useful way in data sets. This suggests that studying more sophisticated unstable homotopy invariants (such as cup products, Massey products, etc.) would be a good direction to pursue. For example, the use of cup products is a key part of Carlsson & Filippenko (2020).
- Within algebraic topology and homotopy theory, a very interesting aspect is the topology of spaces equipped with a reference map to a base space B; this is referred to as parametrized topology. All maps are then required to respect the reference map. The category of spaces over a base contains a much richer set of invariants than in the absolute case (i.e. ordinary topology, without a reference map), where B is a single point. This idea comes up in the study of evasion problems (Carlsson & Filippenko 2020) and can be used to define the idea of data science over a base, or parametrized topological data analysis (Nelson 2020), which appears to be a useful framework for an iterative method of data analysis. The study of unstable invariants in this case is particularly rich, and warrants further attention.
- One is often interested in studying the invariants of a space X which are not necessarily topological in nature but which nevertheless can be thought of as qualitative, for example, the notion of the corners or ends of spaces are examples of this kind of situation. One way to approach such problems is to perform constructions on X so as to produce an associated space which reflects the property one wants to study, and then to use topological methods such as homology to perform the analysis. A powerful example of this philosophy is the work of Simon Donaldson on the topology of smooth 4-manifolds, where he showed that certain moduli spaces attached to smooth 4-manifolds allow one to study the topology of the manifold itself (Donaldson 1984). This kind of approach can be used to investigate various shape distinction problems that are not directly topological in nature.

The goal of this book is to introduce the ideas of topological data analysis to both data scientists and topologists. We have omitted much of the technical material about topology in general, as well as for homology in particular, with the expectation that the reader who has studied the book will be able to go further in the subject as needed. We hope that it will encourage both groups to participate in this exciting intellectual development.

As a matter of convention, we choose to use the terms injective, surjective, and bijective instead of one-to-one, onto, and “one-to-one and onto”.

The authors are very grateful for helpful conversations with many people, including R. Adler, A. Bak, E. Carlsson, J. Carlsson, F. Chazal, J. Curry, V. de Silva, P. Diaconis, H. Edelsbrunner, R. Ghrist, L. Guibas, J. Harer, S. Holmes, M. Lesnick, A. Levine, P. Lum, B. Mann, F. Mémoli, K. Mischaikow, D. Morozov, S. Mukherjee, J. Perea, R. Rabadan, H. Sexton, P. Skraba, G. Singh, R. van de Weijgaert, S. Weinberger, and A. Zomorodian.

We particularly thank A. Blumberg, whose collaboration on early drafts of this book has helped immensely.

Deep thanks also go to the editorial staff at Cambridge University Press, for their patience and all their help.