Mathematics of Data

Hal Schenck

Algebraic Foundations for Applied Topology and Data Analysis

MAX PLANCK INSTITUTE
FOR MATHEMATICS
IN THE SCIENCES

Springer

Mathematics of Data

Volume 1

How to reveal, characterize, and exploit the structure in data? Meeting this central challenge of modern data science requires the development of new mathematical approaches to data analysis, going beyond traditional statistical methods. Fruitful mathematical methods can originate in geometry, topology, algebra, analysis, stochastics, combinatorics, or indeed virtually any field of mathematics. Confronting the challenge of structure in data is already leading to productive new interactions among mathematics, statistics, and computer science, notably in machine learning. We invite novel contributions (research monographs, advanced textbooks, and lecture notes) presenting substantial mathematics that is relevant for data science. Since the methods required to understand data depend on the source and type of the data, we very much welcome contributions comprising significant discussions of the problems presented by particular applications. We also encourage the use of online resources for exercises, software and data sets. Contributions from all mathematical communities that analyze structures in data are welcome. Examples of potential topics include optimization, topological data analysis, compressed sensing, algebraic statistics, information geometry, manifold learning, tensor decomposition, support vector machines, neural networks, and many more.

Hal Schenck

Algebraic Foundations for Applied Topology and Data Analysis

Springer

Hal Schenck
Department of Mathematics and Statistics
Auburn University
Auburn, AL, USA

ISSN 2731-4103 ISSN 2731-4111 (electronic)
Mathematics of Data
ISBN 978-3-031-06663-4 ISBN 978-3-031-06664-1 (eBook)
https://doi.org/10.1007/978-3-031-06664-1

This work was supported by National Science Foundation (2006410) (http://dx.doi.org/10.13039/100000001) and Rosemary Kopel Brown Endowment

Mathematics Subject Classification: 55N31, 18G99

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is a mirror of applied topology and data analysis: it covers a wide range of topics, at levels of sophistication varying from the elementary (matrix algebra) to the esoteric (Grothendieck spectral sequence). My hope is that there is something for everyone, from undergraduates immersed in a first linear algebra class to sophisticates investigating higher dimensional analogs of the barcode. Readers are encouraged to barhop; the goal is to give an intuitive and hands-on introduction to the topics, rather than a punctiliously precise presentation.

The notes grew out of a class taught to statistics graduate students at Auburn University during the COVID summer of 2020. The book reflects that: it is written for a mathematically engaged audience interested in understanding the theoretical underpinnings of topological data analysis. Because the field draws practitioners with a broad range of experience, the book assumes little background at the outset. However, the advanced topics in the latter part of the book require a willingness to tackle technically difficult material.

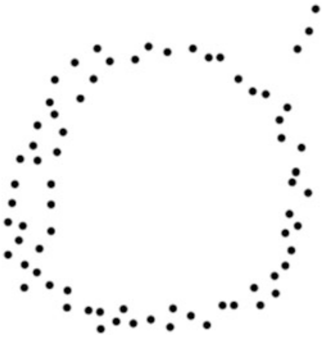
To treat the algebraic foundations of topological data analysis, we need to introduce a fair number of concepts and techniques, so to keep from bogging down, proofs are sometimes sketched or omitted. There are many excellent texts on upper-level algebra and topology where additional details can be found, for example:

Algebra References	**Topology References**
Aluffi [2]	Fulton [76]
Artin [3]	Greenberg–Harper [83]
Eisenbud [70]	Hatcher [91]
Hungerford [93]	Munkres [119]
Lang [102]	Spanier [137]
Rotman [132]	Weibel [149]

Techniques from linear algebra have been essential tools in data analysis from the birth of the field, and so the book kicks off with the basics:

- Least Squares Approximation
- Covariance Matrix and Spread of Data
- Singular Value Decomposition

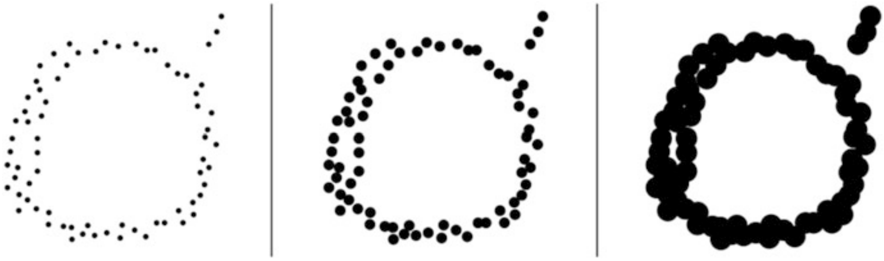
Tools from topology have recently made an impact on data analysis. This text provides the background to understand developments such as *persistent homology*. Suppose we are given point cloud data, that is, a set of points X:



If X was sampled from some object Y, we'd like to use X to infer properties of Y. Persistent homology applies tools of algebraic topology to do this. We start by using X as a *seed* from which to *grow* a family of spaces

$$X_\epsilon = \bigcup_{p \in X} N_\epsilon(p), \text{ where } N_\epsilon(p) \text{ denotes an } \epsilon \text{ ball around } p.$$

As $X_\epsilon \subseteq X_{\epsilon'}$ if $\epsilon \leq \epsilon'$, we have a family of topological spaces and inclusion maps.



As Weinberger notes in [150], persistent homology is a type of Morse theory: there are a finite number of values of ϵ where the topology of X_ϵ changes. Notice that when $\epsilon \gg 0$, X_ϵ is a giant blob; so ϵ is typically restricted to a range $[0, x]$. Topological features which "survive" to the parameter value x are said to be *persistent*; in the example above, the circle S^1 is a persistent feature.

The first three chapters of the book are an algebra-topology boot camp. Chapter 1 provides a brisk review of the tools from linear algebra most relevant for applications, such as webpage ranking. Chapters 2 and 3 cover the results we need from upper-level classes in (respectively) algebra and topology. Applied topology appears in Sect. 3.4, which ties together sheaves, the heat equation, and social media. Some readers may want to skip the first three chapters, and jump in at Chap. 4.

The main techniques appear in Chap. 4, which defines simplicial complexes, simplicial homology, and the Čech & Rips complexes. These concepts are illustrated with a description of the work of de Silva–Ghrist using the Rips complex to analyze sensor networks. Chapter 5 further develops the algebraic topology toolkit, introducing several cohomology theories and highlighting the work of Jiang–Lim–Yao–Ye in applying Hodge theory to voter ranking (the Netflix problem).

We return to algebra in Chap. 6, which is devoted to modules over a principal ideal domain—the structure theorem for modules over a principal ideal domain plays a central role in persistent homology. Chapters 7 and 8 cover, respectively, persistent and multiparameter persistent homology. Chapter 9 is the pièce de résistance (or perhaps the coup de grâce)—a quick and dirty guide to derived functors and spectral sequences. Appendix A illustrates several of the software packages which can be used to perform computations.

There are a number of texts which tackle data analysis from the perspective of pure mathematics. *Elementary Applied Topology* [79] by Rob Ghrist is closest in spirit to these notes; it has a more topological flavor (and wonderfully illuminating illustrations!). Other texts with a similar slant include *Topological Data Analysis with Applications* [30] by Carlsson–Vejdemo-Johanssen, *Computational Topology for Data Analysis* [60] by Dey–Wang, and *Computational Topology* [67] by Edelsbrunner–Harer. At the other end of the spectrum, *Persistence Theory: From Quiver Representations to Data Analysis* [124] by Steve Oudot is intended for more advanced readers; the books of Polterovich–Rosen–Samvelyan–Zhang [126], Rabadan–Blumberg [127], and Robinson [131] focus on applications. Statistical methods are not treated here; they merit a separate volume.

Many folks deserve acknowledgment for their contributions to these notes, starting with my collaborators in the area: Heather Harrington, Nina Otter, and Ulrike Tillmann. A precursor to the summer TDA class was a minicourse I taught at CIMAT, organized by Abraham Martín del Campo, Antonio Rieser, and Carlos Vargas Obieta. The summer TDA class was itself made possible by the Rosemary Kopel Brown endowment, and writing by NSF grant 2048906. Thanks to Hannah Alpert, Ulrich Bauer, Michael Catanzaro, David Cox, Ketai Chen, Carl de Boor, Vin de Silva, Robert Dixon, Jordan Eckert, Ziqin Feng, Oliver Gäfvert, Rob Ghrist, Sean Grate, Anil Hirani, Yang Hui He, Michael Lesnick, Lek-Heng Lim, Dmitriy Morozov, Steve Oudot, Alex Suciu, and Matthew Wright for useful feedback.

Data science is a moving target: one of today's cutting-edge tools may be relegated to the ash bin of history tomorrow. For this reason, the text aims to highlight mathematically important concepts which have *proven* or *potential* utility in applied topology and data analysis. But mathematics is a human endeavor, so it is wise to remember Seneca: "Omnia humana brevia et caduca sunt."

Auburn, AL, USA Hal Schenck
September 2022

Contents