

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

Editor-in-Chief

Jürgen Jost, in den Naturwissenschaften, Max-Planck-Institut für Mathematik,
Leipzig, Germany

Series Editors

Benjamin Gess, Fakultät für Mathematik, Universität Bielefeld, Bielefeld,
Germany

Heather Harrington, Mathematical Institute, University of Oxford, Oxford, UK

Kathryn Hess, Laboratory for Topology and Neuroscience, École Polytechnique
Fédérale de Lausanne, Lausanne, Switzerland

Gitta Kutyniok, Institut für Mathematik, Technische Universität Berlin, Berlin,
Germany

Bernd Sturmfels, in den Naturwissenschaften, Max-Planck-Institut für
Mathematik, Leipzig, Germany

Shmuel Weinberger, Department of Mathematics, University of Chicago, Chicago,
IL, USA

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

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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.

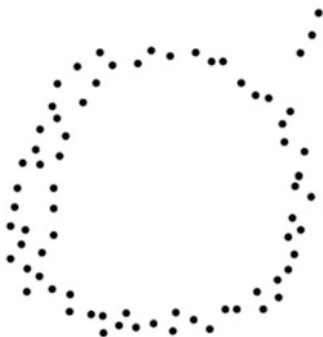
To treat the algebraic foundations of topological data analysis, we need to introduce a fair number of concepts and techniques, so to keep from bogging down, proofs are sometimes sketched or omitted. There are many excellent texts on upper-level algebra and topology where additional details can be found, 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 Boer, 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 Suciú, 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
September 2022

Hal Schenck

Contents

1	Linear Algebra Tools for Data Analysis	1
1.1	Linear Equations, Gaussian Elimination, Matrix Algebra	1
1.2	Vector Spaces, Linear Transformations, Basis and Change of Basis	5
1.2.1	Basis of a Vector Space	6
1.2.2	Linear Transformations	7
1.2.3	Change of Basis	8
1.3	Diagonalization, Webpage Ranking, Data and Covariance	9
1.3.1	Eigenvalues and Eigenvectors	10
1.3.2	Diagonalization	11
1.3.3	Ranking Using Diagonalization	13
1.3.4	Data Application: Diagonalization of the Covariance Matrix	14
1.4	Orthogonality, Least Squares Fitting, Singular Value Decomposition	16
1.4.1	Least Squares	17
1.4.2	Subspaces and Orthogonality	18
1.4.3	Singular Value Decomposition	20
2	Basics of Algebra: Groups, Rings, Modules	23
2.1	Groups, Rings and Homomorphisms	23
2.1.1	Groups	23
2.1.2	Rings	25
2.2	Modules and Operations on Modules	27
2.2.1	Ideals	29
2.2.2	Tensor Product	31
2.2.3	Hom	33
2.3	Localization of Rings and Modules	35
2.4	Noetherian Rings, Hilbert Basis Theorem, Varieties	37
2.4.1	Noetherian Rings	37
2.4.2	Solutions to a Polynomial System: Varieties	38

3	Basics of Topology: Spaces and Sheaves	43
3.1	Topological Spaces	44
3.1.1	Set Theory and Equivalence Relations	44
3.1.2	Definition of a Topology	44
3.1.3	Discrete, Product, and Quotient Topologies	46
3.2	Vector Bundles	48
3.3	Sheaf Theory	51
3.3.1	Presheaves and Sheaves	51
3.3.2	Posets, Direct Limit, and Stalks	53
3.3.3	Morphisms of Sheaves and Exactness	54
3.4	From Graphs to Social Media to Sheaves	55
3.4.1	Spectral Graph Theory	55
3.4.2	Heat Diffusing on a Wire Graph	57
3.4.3	From Graphs to Cellular Sheaves	59
4	Homology I: Simplicial Complexes to Sensor Networks	63
4.1	Simplicial Complexes, Nerve of a Cover	64
4.1.1	The Nerve of a Cover	65
4.2	Simplicial and Singular Homology	66
4.2.1	Singular homology	69
4.3	Snake Lemma and Long Exact Sequence in Homology	71
4.3.1	Maps of complexes, Snake Lemma	71
4.3.2	Chain Homotopy	75
4.4	Mayer–Vietoris, Rips and Čech Complex, Sensor Networks	75
4.4.1	Mayer–Vietoris Sequence	76
4.4.2	Relative Homology	77
4.4.3	Čech Complex and Rips Complex	79
5	Homology II: Cohomology to Ranking Problems	85
5.1	Cohomology: Simplicial, Čech, de Rham Theories	86
5.1.1	Simplicial Cohomology	86
5.1.2	Čech Cohomology	87
5.1.3	de Rham Cohomology	89
5.2	Ranking, the Netflix Problem, and Hodge Theory	93
5.2.1	Hodge Decomposition	93
5.2.2	Application to Ranking	95
5.3	CW Complexes and Cellular Homology	96
5.4	Poincaré and Alexander Duality: Sensor Networks Revisited	98
5.4.1	Statement of Theorems and Examples	99
5.4.2	Alexander Duality: Proof	100
5.4.3	Sensor Networks Revisited	102
5.4.4	Poincaré Duality	105
6	Persistent Algebra: Modules Over a PID	109
6.1	Principal Ideal Domains and Euclidean Domains	110

- 6.2 Rational Canonical Form of a Matrix 112
- 6.3 Linear Transformations, $\mathbb{K}[t]$ -Modules, Jordan Form 113
- 6.4 Structure of Abelian Groups and Persistent Homology 115
 - 6.4.1 \mathbb{Z} -Graded Rings 117
- 7 Persistent Homology** 119
 - 7.1 Barcodes, Persistence Diagrams, Bottleneck Distance 120
 - 7.1.1 History 120
 - 7.1.2 Persistent Homology and the Barcode 121
 - 7.1.3 Computation of Persistent Homology 122
 - 7.1.4 Alpha and Witness Complexes 123
 - 7.1.5 Persistence Diagrams 124
 - 7.1.6 Metrics on Diagrams 125
 - 7.2 Morse Theory 129
 - 7.3 The Stability Theorem 133
 - 7.4 Interleaving and Categories 138
 - 7.4.1 Categories and Functors 138
 - 7.4.2 Interleaving 140
 - 7.4.3 Interleaving Vignette: Merge Trees 144
 - 7.4.4 Zigzag Persistence and Quivers 147
- 8 Multiparameter Persistent Homology** 151
 - 8.1 Definition and Examples 152
 - 8.1.1 Multiparameter Persistence 152
 - 8.2 Graded Algebra, Hilbert Function, Series, Polynomial 155
 - 8.2.1 The Hilbert Function 156
 - 8.2.2 The Hilbert Series 159
 - 8.3 Associated Primes and \mathbb{Z}^n -Graded Modules 161
 - 8.3.1 Geometry of Sheaves 161
 - 8.3.2 Associated Primes and Primary Decomposition 163
 - 8.3.3 Additional Structure in the \mathbb{Z}^n -Graded Setting 167
 - 8.4 Filtrations and Ext 168
- 9 Derived Functors and Spectral Sequences** 173
 - 9.1 Injective and Projective Objects, Resolutions 174
 - 9.1.1 Projective and Injective Objects 174
 - 9.1.2 Resolutions 174
 - 9.2 Derived Functors 175
 - 9.2.1 Categories and Functors 176
 - 9.2.2 Constructing Derived Functors 177
 - 9.2.3 Ext 179
 - 9.2.4 The Global Sections Functor 180
 - 9.2.5 Acyclic Objects 181
 - 9.3 Spectral Sequences 183
 - 9.3.1 Total Complex of Double Complex 183
 - 9.3.2 The Vertical Filtration 184

- 9.3.3 Main Theorem 186
- 9.4 Pas de Deux: Spectral Sequences and Derived Functors 190
 - 9.4.1 Resolution of a Complex 190
 - 9.4.2 Grothendieck Spectral Sequence 192
 - 9.4.3 Comparing Cohomology Theories 194
 - 9.4.4 Cartan–Eilenberg Resolution 196

- A Examples of Software Packages** 201
 - A.1 Covariance and Spread of Data via R 201
 - A.2 Persistent Homology via `scikit-tda` 203
 - A.3 Computational Algebra via `Macaulay2` 207
 - A.4 Multiparameter Persistence via RIVET 209

- Bibliography** 215

- Index** 221