Geometric and Topological Inference

Geometric and topological inference deals with the retrieval of information about a geometric object using only a finite set of possibly noisy sample points. It has connections to manifold learning and provides the mathematical and algorithmic foundations of the rapidly evolving field of topological data analysis. Building on a rigorous treatment of simplicial complexes and distance functions, this self-contained book covers key aspects of the field, from data representation and combinatorial questions to manifold reconstruction and persistent homology. It can serve as a textbook for graduate students or researchers in mathematics, computer science and engineering interested in a geometric approach to data science.

JEAN-DANIEL BOISSONNAT is a Research Director at INRIA, France. His research interests are in Computational Geometry and Topology. He has published several books and more than 180 research papers, and is on the editorial board of the *Journal of the ACM* and of *Discrete and Computational Geometry*. He received the IBM award in Computer Science in 1987, the EADS award in Information Sciences in 2006 and was awarded an advanced grant from the European Research Council in 2014. He has taught at several universities in Paris and at the Collège de France.

FRÉDÉRIC CHAZAL is a Research Director at INRIA, France, where he is heading the DataShape team, a pioneering and world leading group in computational geometry and topological data analysis. His current primary research is on topological data analysis and its connections with statistics and machine learning, and he has authored several reference papers in this domain. He is an associate editor of four international journals and he teaches topological data analysis in various universities and engineering schools in the Paris area.

MARIETTE YVINEC was a researcher at INRIA, France. She is a specialist in the field of shape reconstruction and meshing, and taught master's courses on the subject in various universities in Paris. She coauthored a reference book on computational geometry with Jean-Daniel Boissonnat, and played an active role in the design and development of the software library CGAL.

Cambridge Texts in Applied Mathematics

All titles listed below can be obtained from good booksellers or from Cambridge University Press. For a complete series listing, visit www.cambridge.org/mathematics.

Nonlinear Dispersive Waves
MARK J. ABLOWITZ

Flow, Deformation and Fracture
G. I. BARENBLATT

Hydrodynamic Instabilities
FRANÇOIS CHARRU

The Mathematics of Signal Processing
STEVEN B. DAMELIN & WILLARD MILLER JR.

Introduction to Magnetohydrodynamics (2nd Edition)
P. A. DAVIDSON

An Introduction to Stochastic Dynamics
JINQIAO DUAN

Singularities: Formation, Structure and Propagation
J. EGGERS & M. A. FONTELOS

A Physical Introduction to Suspension Dynamics
ÉLISABETH GUAZZELLI & JEFFREY F. MORRIS

Discrete Systems and Integrability
J. HIETARINTA, N. JOSHI, & F. W. NIJHOFF

An Introduction to Polynomial and Semi-Algebraic Optimization
JEAN BERNARD LASSERRE

An Introduction to Computational Stochastic PDEs
GABRIEL J. LORD, CATHERINE E. POWELL, & TONY SHARDLOW

Numerical Linear Algebra
HOLGER WENDLAND

Geometric and Topological Inference
JEAN-DANIEL BOISSONNAT, FRÉDÉRIC CHAZAL, & MARIETTE YVINEC

Geometric and Topological Inference

JEAN-DANIEL BOISSONNAT
INRIA Sophia Antipolis

FRÉDÉRIC CHAZAL
INRIA Saclay – Ile-de-France

MARIETTE YVINEC
INRIA Sophia Antipolis

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

79 Anson Road, #06–04/06, Singapore 079906

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/9781108419390
DOI: 10.1017/9781108297806

First published 2018

Printed in the United States of America by Sheridan Books, Inc.

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

Library of Congress Cataloging-in-Publication Data
Names: Boissonnat, J.-D. (Jean-Daniel), 1953– author. | Chazal, Frédéric, 1971– author. |
Yvinec, Mariette, 1953– author.
Title: Geometric and topological inference / Jean-Daniel Boissonnat, INRIA
Sophia Antipolis, Frédéric Chazal, Inria Saclay – Ile-de-France,
Mariette Yvinec, INRIA Sophia Antipolis.
Description: New York, NY, USA : Cambridge University Press, 2018. |
Includes bibliographical references and index.
Identifiers: LCCN 2018015875 | ISBN 9781108419390 (Hardback) |
ISBN 9781108410892 (Paperback)
Subjects: LCSH: Shapes–Mathematical models. | Geometric analysis. |
Pattern perception. | Topology.
Classification: LCC QA491 .B5995 2018 | DDC 514/.2–dc23
LC record available at https://lccn.loc.gov/2018015875

ISBN 978-1-108-41939-0 Hardback
ISBN 978-1-108-41089-2 Paperback

Contents

Introduction

Motivation and goals. In many practical situations, geometric objects are only known through a finite set of possibly noisy sample points. A natural question is then to recover the geometry and the topology of the unknown object from this information. The most classical example is probably surface reconstruction, where the points are measured on the surface of a real-world object. A perhaps more surprising example is the study of the large-scale structure formed by the galaxies, which cosmologists believe to be an interconnected network of walls and filaments. In other applications, the shape of interest may be a low-dimensional object embedded in a higher-dimensional space, which is the basic assumption in *manifold learning* [102]. This is for example the case in time series analysis, when the shape of study is the attractor of a dynamical system sampled by a sequence of observations. When these structures are highly nonlinear and have a nontrivial topology, as is often the case, simple dimensionality reduction techniques do not suffice and must be complemented with more geometric and topological techniques.

A lot of research was done in this direction, originating from several sources. A few contributions came from the field of *computational geometry*, where much effort was undertaken to elaborate provably correct surface reconstruction algorithms, under a suitable sampling condition. We refer to [65] for a thorough review of this approach. However, most of this research focused on the case of sampled smooth surfaces in $\mathbb{R}^3$, which is by now fairly well covered. Extending these results to higher-dimensional submanifolds and to nonsmooth objects is one of the objectives of this book. Such an extension requires new data structures to walk around the curse of dimensionality. Handling more general geometric shapes also requires concepts from topology and has provoked an interest in the subject of *computational topology*. Computational topology has recently gained a lot of momentum and has been very successful at providing qualitative invariants and efficient algorithms to compute them. Its application to data analysis led to the rapidly evolving field of *topological data analysis* that provides a general framework to analyze the shape of data and has been applied to various types of data across many fields.

This book. This book intends to cover various aspects of geometric and topological inference, from data representation and combinatorial questions to persistent homology, an adaptation of homology to point cloud data. The aim of this book is not to provide a comprehensive treatment of topological data analysis, but to describe the mathematical and algorithmic foundations of the subject.

Two main concepts will play a central role in this book: simplicial complexes and distance functions. *Simplicial complexes* generalize the notion of triangulation of a surface and are constructed by gluing together simplices: points, line segments, triangles, and their higher-dimensional counterparts. Simplicial complexes can be considered, at the same time, as continuous objects carrying topological and geometric information and as combinatorial data structures that can be efficiently implemented. Simplicial complexes can be used to produce fine meshes leading to faithful approximations well suited to scientific computing purposes, or much coarser approximations, still useful to infer important features of shapes such as their homology or some local geometric properties.

Simplicial complexes have been known and studied for a long time in mathematics but only used in low dimensions due to their high complexity. In this book, we will address the complexity issues by focusing on the inherent, usually unknown, structure in the data which we assume to be of relative low intrinsic dimension. We will put emphasis on output-sensitive algorithms, introduce new simplicial complexes with low complexity, and describe approximation algorithms that scale well with the dimension.

Another central concept in this book is the notion of *distance function*. All the simplicial complexes used in this book encode proximity relationships between the data points. A prominent role is taken by Voronoi diagrams, their dual Delaunay complexes and variants of those, but other simplicial complexes based on distances like the Čech, the Vietoris-Rips, or the witness complexes will also be considered.

This book is divided into four parts.

Part I contains two chapters that present background material on topological spaces and simplicial complexes.

Part II introduces Delaunay complexes and their variants. Since Delaunay complexes are closely related to polytopes, the main combinatorial and algorithmic properties of polytopes are presented first in Chapter 3.

Delaunay complexes, to be introduced in Chapter 4, are defined from Voronoi diagrams, which are natural space partitions induced by the distance function to a sample. Delaunay complexes appear as the underlying basic data structure for manifold reconstruction. The extensions of Voronoi diagrams and Delaunay complexes to weighted distances are also presented together with their relevant applications to kth-nearest neighbor search and Bregman divergences, which are used in information theory, image processing, and statistical analysis.

Although Delaunay triangulations have many beautiful properties, their size depends exponentially on the dimension of the space in the worst case. It is thus important to exhibit realistic assumptions under which the complexity of the Delaunay triangulation does not undergo such a bad behavior. This will be done through the notion of nets. Another issue comes from the fact that, in dimensions greater than 2, Delaunay simplices may have an arbitrarily small volume, even if their vertices are well distributed. Avoiding such bad simplices is a major issue and the importance of thick triangulations has been recognized since the early days of differential topology. They play a central role in numerical simulations to ensure the convergence of numerical methods solving partial differential equations. They also play a central role in the triangulation of manifolds and, in particular, the reconstruction of submanifolds of high-dimensional spaces as shown in Chapter 8. Chapter 5 defines thick triangulations and introduces a random perturbation technique to construct thick Delaunay triangulations in Euclidean space.

Chapter 6 introduces two filtrations of simplicial complexes. Filtrations are nested sequences of subcomplexes that allow to compute persistent homology as described in Chapter 11. We first introduce alpha-complexes and show that they provide natural filtrations of Delaunay and weighted Delaunay complexes. We then introduce witness complexes and their filtrations. The witness complex is a weak version of the Delaunay complex that can be constructed in general metric spaces using only pairwise distances between the points, without a need for coordinates. We will also introduce a filtration of the witness complex.

Part III is devoted to the problem of reconstructing a submanifold $\mathcal{M}$ of $\mathbb{R}^d$ from a finite point sample $P \in \mathcal{M}$. The ultimate goal is to compute a triangulation of $\mathcal{M}$, i.e., a simplicial complex that is homeomorphic to $\mathcal{M}$. This is a demanding quest and, in this part, we will restrict our attention to the case where $\mathcal{M}$ is a smooth submanifold of $\mathbb{R}^d$.

In Chapter 7, we introduce the basic concepts and results, and state a theorem that provides conditions for a simplicial complex $\hat{\mathcal{M}}$ with vertex set $P \subset \mathcal{M}$ to be both a triangulation and a good geometric approximation of $\mathcal{M}$.

Chapter 8 is devoted to the problem of reconstructing submanifolds from point samples. This problem is of primary importance when $\mathcal{M}$ is a surface of $\mathbb{R}^3$ (it is then known as the surface reconstruction problem). It also finds applications in higher dimensions in the context of data analysis where data are considered as points in some Euclidean space, of possibly high dimension. In this chapter, we first exhibit conditions under which the alpha-complex of $P \subset \mathcal{M}$ has the same homotopy type as $\mathcal{M}$, a weaker property than being homeomorphic to $\mathcal{M}$. We then consider the problem of reconstructing a smooth submanifold $\mathcal{M}$ embedded in a space of possibly high dimension d. We then cannot afford to triangulate the ambient space as is being routinely done when working in low dimensions. A way to walk around this difficulty is to assume, as is common practice in data analysis and machine learning, that the intrinsic dimension k of $\mathcal{M}$ is small, even if the dimension of the ambient space may be very large. Chapter 8 takes advantage of this assumption and presents a reconstruction algorithm whose complexity is linear in d and exponential only in k.

The assumptions made in Part III are very demanding: the geometric structures of the data should be smooth submanifolds, the amount of noise in the data should be small and the sampling density should be high. These assumptions may not be satisfied in practical situations. Part IV aims at weakening the assumptions. Chapter 9 studies the stability properties of the sublevel sets of distance functions and provide sampling conditions to infer the underlying geometry and topology of data.

Approximations in Chapter 9 are with respect to the Hausdorff distance. This is a too strong limitation when the data contain outliers that are far away from the underlying structure we want to infer. To overcome this problem, Chapter 10 introduces a new framework where data are no longer considered as points but as distributions of mass or, more precisely probability measures. It is shown that the distance function approach can be extended to this more general framework.

Although Chapters 9 and 10 provide strong results on the topology of the sublevel sets of distance functions, computing and manipulating such sublevel sets is limited in practice to low dimensions. To go beyond these limitations, we restrict our quest to the inference of some topological invariants of the level sets, namely their homology and the associated Betti numbers. Chapter 11 introduces persistent homology and provides tools to robustly infer the homology of sampled shapes.

Efficient implementations of most of the algorithms described in this book can be found in the CGAL library (www.cgal.org/) or in the GUDHI library (http://gudhi.gforge.inria.fr/).

Acknowledgments. This book results from long-standing joint research with several colleagues and collaborators and would not exist without their continuous support and unfailing friendship. We are especially indebted to David Cohen-Steiner, Ramsay Dyer, Vin de Silva, Arijit Ghosh, Marc Glisse, Leonidas Guibas, André Lieutier, Quentin Mérigot, Steve Oudot, and Mathijs Wintraecken who influenced this book in essential ways. This book also includes results obtained together with Clément Maria, Olivier Devillers, Kunal Dutta, and Frank Nielsen. We have been very lucky interacting with them. The book originates from a course that the authors gave at MPRI (*Master Parisien de Recherche en Informatique)*. Many students and colleagues commented on parts of this book. We are particularly grateful to Ramsay Dyer, Arijit Ghosh, and Mathijs Wintraecken for helping shaping Part II, and to Claire Brécheteau for kindly reading Part III and suggesting improvements. We also thank Clément Maria, Mael Rouxel-Labbé, and Mathijs Wintraecken for some of the figures.

The research leading to this book has been partially supported by the Agence Nationale de la Recherche under the project TopData and by the European Research Council under the Advanced Grant GUDHI (Geometric Understanding in Higher Dimensions).