Abel Symposia 15

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

# Topological Data Analysis

## The Abel Symposium 2018

Springer

ABEL SYMPOSIA

Edited by the Norwegian Mathematical Society

More information about this series at http://www.springer.com/series/7462

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

# Topological Data Analysis

## The Abel Symposium 2018

ABEL
PRISEN

Springer

*Editors*
Nils A. Baas
Department of Mathematical Sciences
NTNU
Trondheim, Norway

Gunnar E. Carlsson
Department of Mathematics
Stanford University
Stanford, CA, USA

Gereon Quick
Department of Mathematical Sciences
NTNU
Trondheim, Norway

Markus Szymik
Department of Mathematical Sciences
NTNU
Trondheim, Norway

Marius Thaule
Department of Mathematicsal Sciences
NTNU
Trondheim, Norway

ISSN 2193-2808 ISSN 2197-8549 (electronic)
Abel Symposia
ISBN 978-3-030-43407-6 ISBN 978-3-030-43408-3 (eBook)
https://doi.org/10.1007/978-3-030-43408-3

Mathematics Subject Classification: 55-XX

# Series Foreword

The Norwegian government established the Abel Prize in mathematics in 2002, and the first prize was awarded in 2003. In addition to honoring the great Norwegian mathematician Niels Henrik Abel by awarding an international prize for outstanding scientific work in the field of mathematics, the prize shall contribute toward raising the status of mathematics in society and stimulate the interest for science among school children and students. In keeping with this objective, the Niels Henrik Abel Board has decided to finance annual Abel Symposia. The topic of the symposia may be selected broadly in the area of pure and applied mathematics. The symposia should be at the highest international level and serve to build bridges between the national and international research communities. The Norwegian Mathematical Society is responsible for the events. It has also been decided that the contributions from these symposia should be presented in a series of proceedings, and Springer Verlag has enthusiastically agreed to publish the series. The Niels Henrik Abel Board is confident that the series will be a valuable contribution to the mathematical literature.

Chair of the Niels Henrik Abel Board — John Grue

# Preface

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 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, 5–7, 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.

Trondheim, Norway — Nils A. Baas
Stanford, California, CA, USA — Gunnar E. Carlsson
Trondheim, Norway — Gereon Quick
Trondheim, Norway — Markus Szymik
Trondheim, Norway — Marius Thaule
October 2019

## 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. 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)
15. Edelsbrunner H., Letscher D., Zomorodian A.: Topological persistence and simplification. Discrete and Computational Geometry **28**, 511–533 (2002)
16. Cagliari F., B. Di Fabio B., Ferri M.: One-dimensional reduction of multidimensional persistent homology. Proc. Amer. Mat. Soc. **138**, 3003–3017 (2010)
17. 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
18. 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
19. Kanari L., Dlotko P., Scolamiero M., Levi R., Shillcock J.C., Hess K., Markram H.: A topological representation of branching morphologies. Neuroinformatics (2017)
20. Lee J-K., et al: Spatiotemporal genomic architecture informs precision oncology in glioblastoma. Nature Genetics **49**, 594–599 (2017)
21. MacPherson R., Schweinhart B.: Measuring shape with topology. J. Math. Phys. **53** (2012)
22. 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
23. 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
24. 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)
25. 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)
26. Robinson M.: Topological Signal Processing. Springer Verlag (2014)
27. Rybakken E., Baas N., Dunn B.: Decoding of neural data using cohomological features extraction. Neural Computation **31**, 68–93 (2019)
28. 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)
29. 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)
30. 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)
31. Zomorodian A., Carlsson G.: Computing persistent homology, Discrete and Computational Geometry. **33**, 249–274 (2005)

# Contents

# A Fractal Dimension for Measures via Persistent Homology


**Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Joshua Mirth, Rachel Neville, Chris Peterson, and Clayton Shonkwiler**

**Abstract** We use persistent homology in order to define a family of fractal dimensions, denoted $\dim_{\mathrm{PH}}^{i}(\mu)$ for each homological dimension $i \geq 0$, assigned to a probability measure $\mu$ on a metric space. The case of zero-dimensional homology ($i = 0$) relates to work by Steele (Ann Probab 16(4): 1767–1787, 1988) studying the total length of a minimal spanning tree on a random sampling of points. Indeed, if $\mu$ is supported on a compact subset of Euclidean space $\mathbb{R}^m$ for $m \geq 2$, then Steele's work implies that $\dim_{\mathrm{PH}}^{0}(\mu) = m$ if the absolutely continuous part of $\mu$ has positive mass, and otherwise $\dim_{\mathrm{PH}}^{0}(\mu) < m$. Experiments suggest that similar results may be true for higher-dimensional homology $0 < i < m$, though this is an open question. Our fractal dimension is defined by considering a limit, as the number of points $n$ goes to infinity, of the total sum of the $i$-dimensional persistent homology interval lengths for $n$ random points selected from $\mu$ in an i.i.d. fashion. To some measures $\mu$, we are able to assign a finer invariant, a curve measuring the limiting distribution of persistent homology interval lengths as the number of points goes to infinity. We prove this limiting curve exists in the case of zero-dimensional homology when $\mu$ is the uniform distribution over the unit interval, and

This work was completed while Elin Farnell was a research scientist in the Department of Mathematics at Colorado State University.

H. Adams (✉) · M. Aminian · M. Kirby · J. Mirth · C. Peterson · C. Shonkwiler
Colorado State University, Fort Collins, CO, USA
e-mail: adams@math.colostate.edu; Manuchehr.Aminian@colostate.edu; kirby@math.colosate.edu; mirth@math.colostate.edu; peterson@math.colostate.edu; clayton@math.colostate.edu

E. Farnell
Amazon, Seattle, WA, USA
e-mail: efarnell@amazon.com

R. Neville
University of Arizona, Fort Collins, CO, USA
e-mail: raneville@math.arizona.edu



N. A. Baas et al. (eds.), *Topological Data Analysis*, Abel Symposia 15,
https://doi.org/10.1007/978-3-030-43408-3_1

conjecture that it exists when $\mu$ is the rescaled probability measure for a compact set in Euclidean space with positive Lebesgue measure.

## 1 Introduction

Let $X$ be a metric space equipped with a probability measure $\mu$. While fractal dimensions are most classically defined for a space, there are a variety of fractal dimension definitions for a measure, including the Hausdorff or packing dimension of a measure [24, 30, 54]. In this paper we use persistent homology to define a fractal dimension $\dim_{\mathrm{PH}}^{i}(\mu)$ associated to a measure $\mu$ for each homological dimension $i \geq 0$. Roughly speaking, $\dim_{\mathrm{PH}}^{i}(\mu)$ is determined by how the lengths of the persistent homology intervals for a random sample, $X_n$, of $n$ points from $X$ vary as $n$ tends to infinity.

Our definition should be thought of as a generalization, to higher homological dimensions, of fractal dimensions related to minimal spanning trees, as studied, for example, in [63]. Indeed, the lengths of the zero-dimensional (reduced) persistent homology intervals corresponding to the Vietoris–Rips complex of a sample $X_n$ are equal to the lengths of the edges in a minimal spanning tree with $X_n$ as the set of vertices. In particular, if $X$ is a subset of Euclidean space $\mathbb{R}^m$ with $m \geq 2$, then [63, Theorem 1] by Steele implies that $\dim_{\mathrm{PH}}^{0}(\mu) \leq m$, with equality when the absolutely continuous part of $\mu$ has positive mass (Proposition 1). Independent generalizations of Steele's work to higher homological dimensions are considered in [26, 61, 62].

To some metric spaces $X$ equipped with a measure $\mu$ we are able to assign a finer invariant that contains more information than just the fractal dimension. Consider the set of the lengths of all intervals in the $i$-dimensional persistent homology for $X_n$. Experiments suggest that when probability measure $\mu$ is absolutely continuous with respect to the Lebesgue measure on $X \subseteq \mathbb{R}^m$, the scaled set of interval lengths in each homological dimension $i$ converges distribution-wise to some fixed probability distribution (depending on $\mu$ and $i$). This is easy to prove in the simple case of zero-dimensional homology when $\mu$ is the uniform distribution over the unit interval, in which case we can also derive a formula for the limiting distribution. Experiments suggest that when $\mu$ is the rescaled probability measure corresponding to a compact set $X \subseteq \mathbb{R}^m$ of positive Lebesgue measure, then a limiting rescaled distribution exists that depends only on $m$, $i$, and the volume of $\mu$ (see Conjecture 2). We would be interested to know the formulas for the limiting distributions with higher Euclidean and homological dimensions.

Whereas Steele in [63] studies minimal spanning trees on random subsets of a space, Kozma et al. in [42] study minimal spanning trees built on extremal subsets. Indeed, they define a fractal dimension for a metric space $X$ as the infimum, over all powers $d$, such that for any minimal spanning tree $T$ on a finite number of points in $X$, the sum of the edge lengths in $T$ each raised to the power $d$ is bounded. They relate this extremal minimal spanning tree dimension to the box counting dimension. Their work is generalized to higher homological dimensions by

Schweinhart [60]. By contrast, we instead generalize Steele's work [63] on measures to higher homological dimensions. Three differences between [42, 60] and our work are the following.

- The former references define a fractal dimension for metric spaces, whereas we define a fractal dimension for measures.
- The fractal dimension in [42, 60] is defined using extremal subsets, whereas we define our fractal dimension using random subsets.
- We can estimate our fractal dimension computationally using log-log plots as in Sect. 5, whereas we do not know a computational technique for estimating the fractal dimensions in [42, 60].

After describing related work in Sect. 2, we give preliminaries on fractal dimensions and on persistent homology in Sect. 3. We present the definition of our fractal dimension and prove some basic properties in Sect. 4. We demonstrate example experimental computations in Sect. 5; our code is publicly available at https://github.com/CSU-PHdimension/PHdimension. Section 6 describes how limiting distributions, when they exist, form a finer invariant. Sects. 7 and 8 discuss the computational details involved in sampling from certain fractals and estimating asymptotic behavior, respectively. Finally we present our conclusion in Sect. 9. One of the main goals of this paper is to pose questions and conjectures, which are shared throughout.

## 2 Related Work

### *2.1 Minimal Spanning Trees*

The paper [63] studies the total length of a minimal spanning tree for random subsets of Euclidean space. Let $X_n$ be a random sample of points from a compact subset of $\mathbb{R}^d$ according to some probability distribution. Let $M_n$ be the sum of all the edge lengths of a minimal spanning tree on vertex set $X_n$. Then for $d \geq 2$, Theorem 1 of [63] says that

$$M_n \sim C n^{(d-1)/d} \quad \text{as} \quad n \to \infty, \tag{1.1}$$

where the relation $\sim$ denotes asymptotic convergence, with the ratio of the terms approaching one in the specified limit. Here, $C$ is a fixed constant depending on $d$ and on the *volume* of the absolutely continuous part of the probability distribution.[1] There has been a wide variety of related work, including for example [5–7, 38, 64–67]. See [41] for a version of the central limit theorem in this context. The papers [51, 52] study the length of the longest edge in the minimal spanning tree

[1] If the compact subset has Hausdorff dimension less than $d$, then [63] implies $C = 0$.

for points sampled uniformly at random from the unit square, or from a torus of dimension at least two. By contrast, [42] studies Euclidean minimal spanning trees built on extremal finite subsets, as opposed to random subsets.

## 2.2 Umbrella Theorems for Euclidean Functionals

As Yukich explains in his book [72], there are a wide variety of Euclidean functionals, such as the length of the minimal spanning tree, the length of the traveling salesperson tour, and the length of the minimal matching, which all have scaling asymptotics analogous to (1.1). To prove such results, one needs to show that the Euclidean functional of interest satisfies translation invariance, subadditivity, superadditivity, and continuity, as in [21, Page 4]. Superadditivity does not always hold, for example it does not hold for the minimal spanning tree length functional, but there is a related "boundary minimal spanning tree functional" that does satisfy superadditivity. Furthermore, the boundary functional has the same asymptotics as the original functional, which is enough to prove scaling results. It is intriguing to ask if these techniques will work for functionals defined using higher-dimensional homology.

## 2.3 Random Geometric Graphs

In this paper we consider simplicial complexes (say Vietoris–Rips or Čech) with randomly sampled points as the vertex set. The 1-skeleta of these simplicial complexes are random geometric graphs. We recommend the book [50] by Penrose as an introduction to random geometric graphs; related families of random graphs are also considered in [53]. Random geometric graphs are often studied when the scale parameter $r(n)$ is a function of the number of vertices $n$, with $r(n)$ tending to zero as $n$ goes to infinity. Instead, in this paper we are more interested in the behavior over all scale parameters simultaneously. From a slightly different perspective, the paper [40] studies the expected Euler characteristic of the union of randomly sampled balls (potentially of varying radii) in the plane.

## 2.4 Persistent Homology

Vanessa Robins' thesis [58] contains many related ideas; we describe one such example here. Given a set $X \subseteq \mathbb{R}^m$ and a scale parameter $\varepsilon \geq 0$, let

$$X_\varepsilon = \{y \in \mathbb{R}^m \mid \text{there exists some } x \in X \text{ with } d(y, x) \leq \varepsilon\}$$

denote the *ε-offset of* $X$. The $\varepsilon$-offset of $X$ is equivalently the union of all closed $\varepsilon$ balls centered at points in $X$. Furthermore, let $C(X_\varepsilon) \in \mathbb{N}$ denote the number of connected components of $X_\varepsilon$. In Chapter 5, Robins shows that for a generalized Cantor set $X$ in $\mathbb{R}$ with Lebesgue measure 0, the box-counting dimension of $X$ is equal to the limit

$$\lim_{\varepsilon \to 0} \frac{\log(C(X_\varepsilon))}{\log(1/\varepsilon)}.$$

Here Robins considers the entire Cantor set, whereas we study random subsets thereof.

The paper [46], which heavily influenced our work, introduces a fractal dimension defined using persistent homology. This fractal dimension depends on thickenings of the entire metric space $X$, as opposed to random or extremal subsets thereof. As a consequence, the computed dimension of some fractal shapes (such as the Cantor set cross the interval) disagrees significantly with the Hausdorff or box-counting dimension.

Schweinhart's paper [60] takes a slightly different approach from ours, considering extremal (as opposed to random) subsets. After fixing a homological dimension $i$, Schweinhart assigns a fractal dimension to each metric space $X$ equal to the infimum over all powers $d$ such that for any finite subset $X' \subseteq X$, the sum of the $i$-dimensional persistent homology bar lengths for $X'$, each raised to the power $d$, is bounded. For low-dimensional metric spaces Schweinhart relates this dimension to the box counting dimension.

More recently, Divol and Polonik [26] obtain generalizations of [63, 72] to higher homological dimensions in the case when $X$ is a cube. Related results are obtained in [62] when $X$ is a ball or sphere, and afterwards in [61] when points are sampled according to an Ahlfors regular measure.

There is a growing literature on the topology of random geometric simplicial complexes, including in particular the homology of Vietoris–Rips and Čech complexes built on top of random points in Euclidean space [3, 13, 39]. The paper [14] shows that for $n$ points sampled from the unit cube $[0, 1]^d$ with $d \geq 2$, the maximally persistent cycle in dimension $1 \leq k \leq d-1$ has persistence of order $\Theta((\frac{\log n}{\log \log n})^{1/k})$, where the asymptotic notation big Theta means both big O and big Omega. The homology of Gaussian random fields is studied in [4], which gives the expected $k$-dimensional Betti numbers in the limit as the number of points increases to infinity, and also in [12]. The paper [29] studies the number of simplices and critical simplices in the alpha and Delaunay complexes of Euclidean point sets sampled according to a Poisson process. An open problem about the birth and death times of the points in a persistence diagram coming from sublevelsets of a Gaussian random field is stated in Problem 1 of [28]. The paper [18] shows that the expected persistence diagram,from a wide class of random point clouds, has a density with respect to the Lebesgue measure

The paper [15] explores what attributes of an algebraic variety can be estimated from a random sample, such as the variety's dimension, degree, number of irreducible components, and defining polynomials; one of their estimates of dimension is inspired by our work.

In an experiment in [1], persistence diagrams are produced from random subsets of a variety of synthetic metric space classes. Machine learning tools, with these persistence diagrams as input, are then used to classify the metric spaces corresponding to each random subset. The authors obtain high classification rates between the different metric spaces. It is likely that the discriminating power is based not only on the underlying homotopy types of the shape classes, but also on the shapes' dimensions as detected by persistent homology.

## 3 Preliminaries

This section contains background material and notation on fractal dimensions and persistent homology.

### *3.1 Fractal Dimensions*

The concept of fractal dimension was introduced by Hausdorff to describe spaces like the Cantor set, and it later found extensive application in the study of dynamical systems. The attracting sets of simple a dynamical system is often a submanifold, with an obvious dimension, but in non-linear and chaotic dynamical systems the attracting set may not be a manifold. The Cantor set, defined by removing the middle third from the interval [0, 1], and then recursing on the remaining pieces, is a typical example. It has the same cardinality as $\mathbb{R}$, but it is nowhere-dense, meaning it at no point resembles a line. The typical fractal dimension of the Cantor set is $\log_3(2)$. Intuitively, the Cantor set has "too many" points to have dimension zero, but also should not have dimension one.

We speak of fractal dimensions in the plural because there are many different definitions. In particular, fractal dimensions can be divided into two classes, which have been called "metric" and "probabilistic" [31]. The former describe only the geometry of a metric space. Two widely-known definitions of this type, which often agree on well-behaved fractals, but are not in general equal, are the box-counting and Hausdorff dimensions. For an inviting introduction to fractal dimensions see [30]. Dimensions of the latter type take into account both the geometry of a given set and a probability distribution supported on that set—originally the "natural measure" of the attractor given by the associated dynamical system, but in principle any probability distribution can be used. The information dimension is the best known example of this type. For detailed comparisons, see [32]. Our persistent homology fractal dimension, Definition 6, is of the latter type.

For completeness, we exhibit some of the common definitions of fractal dimension. The primary definition for sets is given by the Hausdorff dimension [33].

**Definition 1** Let $S$ be a subset of a metric space $X$, let $d \in [0, \infty)$, and let $\delta > 0$. The *Hausdorff measure* of $S$ is

$$H_d(S) = \inf_{\delta} \left( \inf \left\{ \sum_{j=1}^{\infty} \operatorname{diam}(B_j)^d \mid S \subseteq \bigcup_{j=1}^{\infty} B_j \text{ and } \operatorname{diam}(B_j) \leq \delta \right\} \right),$$

where the inner infimum is over all coverings of $S$ by balls $B_j$ of diameter at most $\delta$. The *Hausdorff dimension* of $S$ is

$$\dim_H(S) = \inf_d \{H_d(S) = 0.\}$$

The Hausdorff dimension of the Cantor set, for example, is $\log_3(2)$.

In practice it is difficult to compute the Hausdorff dimension of an arbitrary set, which has led to a number of alternative fractal dimension definitions in the literature. These dimensions tend to agree on well-behaved fractals, such as the Cantor set, but they need not coincide in general. Two worth mentioning are the box-counting dimension, which is relatively simple to define, and the correlation dimension.

**Definition 2** Let $S \subseteq X$ a metric space, and let $N_\varepsilon$ denote the infimum of the number of closed balls of radius $\epsilon$ required to cover $S$. Then the *box-counting dimension* of $S$ is

$$\dim_B(S) = \lim_{\varepsilon \to 0} \frac{\log(N_\varepsilon)}{\log(1/\varepsilon)},$$

provided this limit exists. Replacing the limit with a lim sup gives the *upper* box-counting dimension, and a lim inf gives the *lower* box-counting dimension.

The box-counting definition is unchanged if $N_\epsilon$ is instead defined by taking the number of open balls of radius $\varepsilon$, or the number of sets of diameter at most $\varepsilon$, or (for $S$ a subset of $\mathbb{R}^n$) the number of cubes of side-length $\varepsilon$ [70, Definition 7.8], [30, Equivalent Definitions 2.1]. It can be shown that $\dim_B(S) \geq \dim_H(S)$. This inequality can be strict; for example if $S = \mathbb{Q} \cap [0, 1]$ is the set of all rational numbers between zero and one, then $\dim_H(S) = 0 < 1 = \dim_B(S)$ [30, Chapter 3].

In Sect. 4 we introduce a fractal dimension based on persistent homology which shares key similarities with the Hausdorff and box-counting dimensions. It can also be easily estimated via log-log plots, and it is defined for arbitrary metric spaces (though our examples will tend to be subsets of Euclidean space). A key difference, however, will be that ours is a fractal dimension for measures, rather than for subsets.

There are a variety of classical notions of a fractal dimension for a measure, including the Hausdorff, packing, and correlation dimensions of a measure [24, 30, 54]. We give the definitions of two of these.

**Definition 3 ((13.16) of [30])** The *Hausdorff dimension* of a measure $\mu$ with total mass one is defined as

$$\dim_H(\mu) = \inf\{\dim_H(S) \mid S \text{ is a Borel subset with } \mu(S) > 0\}.$$

We have $\dim_H(\mu) \leq \dim_H(\text{supp}(\mu))$, and it is possible for this inequality to be strict [30, Exercise 3.10].[2] We also give the example of the correlation dimension of a measure.

**Definition 4** Let $X$ be a subset of $\mathbb{R}^m$ equipped with a measure $\mu$, and let $X_n$ be a random sample of $n$ points from $X$. Let $\theta\colon \mathbb{R} \to \mathbb{R}$ denote the Heaviside step function, meaning $\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x \geq 0$. The *correlation integral* of $\mu$ is defined (for example in [35, 69]) to be

$$C(r) = \lim_{n\to\infty} \frac{1}{n^2} \sum_{\substack{x,x' \in X_n \\ x \neq x'}} \theta\left(r - \|x - x'\|\right).$$

It can be shown that $C(r) \propto r^{\nu}$, and the exponent $\nu$ is defined to be the *correlation dimension* of $\mu$.

In [35, 36] it is shown that the correlation dimension gives a lower bound on the Hausdorff dimension of a measure. The correlation dimension can be easily estimated from a log-log plot, similar to the methods we use in Sect. 5. A different definition of the correlation definition is given and studied in [23, 47]. The correlation dimension is a particular example of the family of *Rènyi dimensions*, which also includes the *information dimension* as a particular case [56, 57]. A collection of possible axioms that one might like to have such a fractal dimension satisfy is given in [47].

## 3.2 Persistent Homology

The field of applied and computational topology has grown rapidly in recent years, with the topic of persistent homology gaining particular prominence. Persistent homology has enjoyed a wealth of meaningful applications to areas such as image analysis, chemistry, natural language processing, and neuroscience, to name just a

[2] See also [31] for an example of a measure whose *information dimension* is less than the Hausdorff dimension of its support.

few examples [2, 10, 20, 25, 44, 45, 71, 73]. The strength of persistent homology lies in its ability to characterize important features in data across multiple scales. Roughly speaking, homology provides the ability to count the number of independent $k$-dimensional holes in a space, and persistent homology provides a means of tracking such features as the scale increases. We provide a brief introduction to persistent homology in this preliminaries section, but we point the interested reader to [8, 27, 37] for thorough introductions to homology, and to [16, 22, 34] for excellent expository articles on persistent homology.

Geometric complexes, which are at the heart of the work in this paper, associate to a set of data points a simplicial complex—a combinatorial space that serves as a model for an underlying topological space from which the data has been sampled. The building blocks of simplicial complexes are called simplices, which include vertices as 0-simplices, edges as 1-simplices, triangles as 2-simplices, tetrahedra as 3-simplices, and their higher-dimensional analogues as $k$-simplices for larger values of $k$. An important example of a simplicial complex is the Vietoris–Rips complex.

**Definition 5** Let $X$ be a set of points in a metric space and let $r \geq 0$ be a scale parameter. We define the Vietoris–Rips simplicial complex $\mathrm{VR}(X; r)$ to have as its $k$-simplices those collections of $k + 1$ points in $X$ that have diameter at most $r$.

In constructing the Vietoris–Rips simplicial complex we translate our collection of points in $X$ into a higher-dimensional complex that models topological features of the data. See Fig. 1 for an example of a Vietoris–Rips complex constructed from a set of data points, and see [27] for an extended discussion.

It is readily observed that for various data sets, there is not necessarily an ideal choice of the scale parameter so that the associated Vietoris–Rips complex captures the desired features in the data. The perspective behind persistence is to instead allow the scale parameter to increase and to observe the corresponding appearance and disappearance of topological features. To be more precise, each hole appears at a certain scale and disappears at a larger scale. Those holes that persist across a wide range of scales often reflect topological features in the shape underlying the data, whereas the holes that do not persist for long are often considered to be noise.

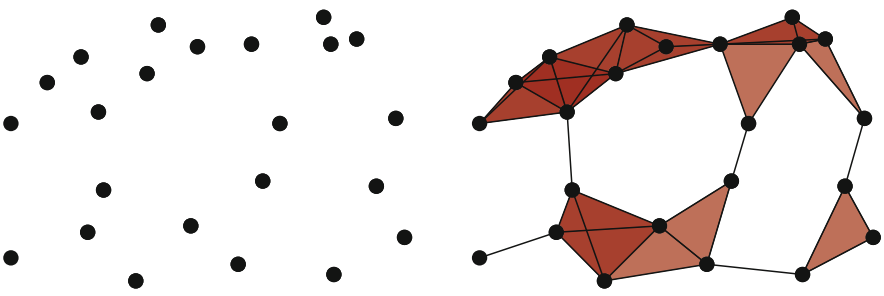

**Fig. 1** An example of a set of data points in $\mathbb{R}^m$ with an associated Vietoris–Rips complex at a fixed scale

However, in the context of this paper (estimating fractal dimensions), the holes that do not persist are perhaps better described as measuring the local geometry present in a random finite sample.

For a fixed set of points, we note that as scale increases, simplices can only be added and cannot be removed. Thus, for $r_0 < r_1 < r_2 < \cdots$, we obtain a filtration of Vietoris–Rips complexes

$$\mathrm{VR}(X; r_0) \subseteq \mathrm{VR}(X; r_1) \subseteq \mathrm{VR}(X; r_2) \subseteq \cdots .$$

The associated inclusion maps induce linear maps between the corresponding homology groups $H_k(\mathrm{VR}(X; r_i))$, which are algebraic structures whose ranks count the number of independent $k$-dimensional holes in the Vietoris–Rips complex. A technical remark is that homology depends on the choice of a group of coefficients; it is simplest to use field coefficients (for example $\mathbb{R}$, $\mathbb{Q}$, or $\mathbb{Z}/p\mathbb{Z}$ for $p$ prime), in which case the homology groups are furthermore vector spaces. The corresponding collection of vector spaces and linear maps is called a *persistent homology module*.

A useful tool for visualizing and extracting meaning from persistent homology is a barcode. The basic idea is that each generator of persistent homology can be represented by an interval, whose start and end times are the *birth* and *death* scales of a homological feature in the data. These intervals can be arranged as a barcode graph in which the $x$-axis corresponds to the scale parameter. See Fig. 2 for an example. If $Y$ is a finite metric space, then we let $\mathrm{PH}^i(Y)$ denote the corresponding collection of $i$-dimensional persistent homology intervals.

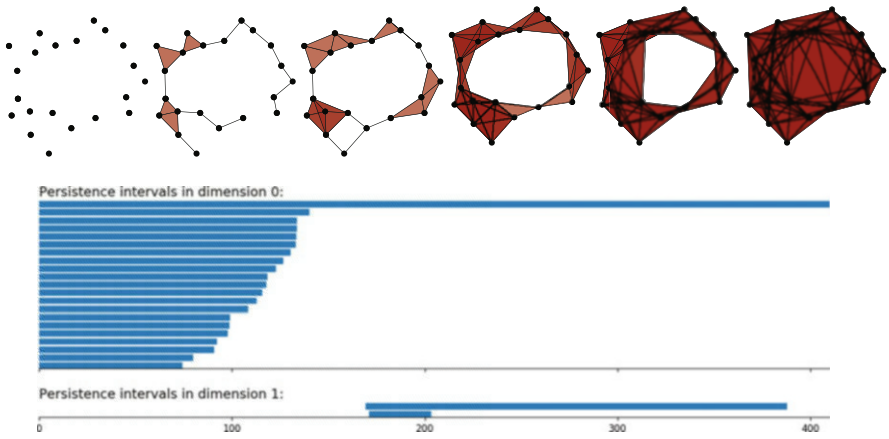


**Fig. 2** An example of Vietoris–Rips complexes at increasing scales, along with associated persistent homology intervals. The zero-dimensional persistent homology intervals shows how 21 connected components merge into a single connected component as the scale increases. The one-dimensional persistent homology intervals show two one-dimensional holes, one short-lived and the other long-lived

Zero-dimensional barcodes always produce one infinite interval, as in Fig. 2, which are problematic for our purposes. Therefore, in the remainder of this paper we will always use reduced homology, which has the effect of simply eliminating the infinite interval from the zero-dimensional barcode while leaving everything else unchanged. As a consequence, there will never be any infinite intervals in the persistent homology of a Vietoris–Rips simplicial complex, even in homological dimension zero.

*Remark 1* It is well-known (see for example [58]) and easy to verify that for any finite metric space $X$, the lengths of the zero-dimensional (reduced) persistent homology intervals of the Vietoris–Rips complex of $X$ correspond exactly to the lengths of the edges in a minimal spanning tree with vertex set $X$.

## 4 Definition of the Persistent Homology Fractal Dimension for Measures

Let $X$ be a metric space equipped with a probability measure $\mu$, and let $X_n \subseteq X$ be a random sample of $n$ points from $X$ distributed independently and identically according to $\mu$. Build a filtered simplicial complex $K$ on top of vertex set $X_n$, for example a Vietoris–Rips complex $\mathrm{VR}(X; r)$ (Definition 5), an intrinsic Čech complex $\check{\mathrm{C}}(X, X; r)$, or an ambient Čech complex $\check{\mathrm{C}}(X, \mathbb{R}^m; r)$ if $X$ is a subset of $\mathbb{R}^m$ [17]. Denote the $i$-dimensional persistent homology of this filtered simplicial complex by $\mathrm{PH}^i(X_n)$. This persistent homology barcode decomposes as a direct sum of interval summands; we let $L^i(X_n)$ be the sum of the lengths of the intervals in $\mathrm{PH}^i(X_n)$. In the case of homological dimension zero, the sum $L^0(X_n)$ is simply the sum of all the edge lengths in a minimal spanning tree with $X_n$ as its vertex set (since we are using reduced homology).

**Definition 6 (Persistent Homology Fractal Dimension)** Let $X$ be a metric space equipped with a probability measure $\mu$, let $X_n \subseteq X$ be a random sample of $n$ points from $X$ distributed according to $\mu$, and let $L^i(X_n)$ be the sum of the lengths of the intervals in the $i$-dimensional persistent homology for $X_n$. We define the *$i$-dimensional persistent homology fractal dimension of* $\mu$ to be

$$\dim^i_{\mathrm{PH}}(\mu) = \inf_{d>0} \Big\{ d \;\Big|\; \exists \text{ constant } C(i, \mu, d) \text{ such that } L^i(X_n) \leq C n^{(d-1)/d} \text{ with probability one as } n \to \infty \Big\}.$$

The constant $C$ can depend on $i$, $\mu$, and $d$. Here "$L^i(X_n) \leq C n^{(d-1)/d}$ with probability one as $n \to \infty$" means that we have $\lim_{n\to\infty} \mathbb{P}[L^i(X_n) \leq C n^{(d-1)/d}] = 1$. This dimension may depend on the choices of filtered simplicial complex (say Vietoris–Rips or Čech), and on the choice of field coefficients for homology computations; for now those choices are suppressed from the definition.

**Proposition 1** *Let $\mu$ be a measure on $X \subseteq \mathbb{R}^m$ with $m \geq 2$. Then $\dim_{\mathrm{PH}}^0(\mu) \leq m$, with equality if the absolutely continuous part of $\mu$ has positive mass.*

***Proof*** By Theorem 2 of [63], we have that $\lim_{n\to\infty} n^{-(m-1)/m} L^0(X_n) = c \int_{\mathbb{R}^m} f(x)^{(m-1)/m}\, dx$, where $c$ is a constant depending on $m$, and where $f$ is the absolutely continuous part of $\mu$. To see that $\dim_{\mathrm{PH}}^0(\mu) \leq m$, note that

$$L^0(X_n) \leq \left( c \int_{\mathbb{R}^m} f(x)^{(m-1)/m}\, dx + \varepsilon \right) n^{(m-1)/m}$$

with probability one as $n \to \infty$ for any $\varepsilon > 0$. □

We conjecture that the $i$-dimensional persistent homology of compact subsets of $\mathbb{R}^m$ have the same scaling properties as the functionals in [63, 72].

*Conjecture 1* Let $\mu$ be a probability measure on a compact set $X \subseteq \mathbb{R}^m$ with $m \geq 2$, and let $\mu$ be absolutely continuous with respect to the Lebesgue measure. Then for all $0 \leq i < m$, there is a constant $C \geq 0$ (depending on $\mu$, $m$, and $i$) such that $L^i(X_n) = Cn^{(m-1)/m}$ with probability one as $n \to \infty$.

Let $\mu$ be a probability measure with compact support that is absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^m$ for $m \geq 2$. Note that Conjecture 1 would imply that the persistent homology fractal dimension of $\mu$ is equal to $m$. The tools of subadditivity and superadditivity behind the umbrella theorems for Euclidean functionals, as described in [72] and Sect. 2.2, may be helpful towards proving this conjecture. In some limited cases, for example when $X$ is a cube or ball, or when $\mu$ is Ahlfors regular, then Conjecture 1 is closely related to [26, 61, 62].

One could alternatively define birth-time or death-time fractal dimensions by replacing $L^i(X_n)$ with the sum of the birth times, or alternatively the sum of the death times, in the persistent homology barcodes $\mathrm{PH}^i(X_n)$.

## 5 Experiments

A feature of Definition 6 is that we can use it to estimate the persistent homology fractal dimension of a measure $\mu$. Indeed, suppose we can sample from $X$ according to the probability distribution $\mu$. We can therefore sample collections of points $X_n$ of size $n$, compute the statistic $L^i(X_n)$, and then plot the results in a log-log fashion as $n$ increases. In the limit as $n$ goes to infinity, we expect the plotted points to be well-modeled by a line of slope $\frac{d-1}{d}$, where $d$ is the $i$-dimensional persistent homology fractal dimension of $\mu$. In many of the experiments in this section, the measures $\mu$ are simple enough (or self-similar enough) that we would expect the persistent homology fractal dimension of $\mu$ to be equal to the Hausdorff dimension of $\mu$.

In our computational experiments, we have used the persistent homology software packages Ripser [9], Javaplex [68], and code from Duke (see the acknowledgements). For the case of zero-dimensional homology, we can alternatively use well-known algorithms for computing minimal spanning trees, such as Kruskal's algorithm or Prim's algorithm [43, 55]. We estimate the slope of our log-log plots (of $L^i(X_n)$ as a function of $n$) using both a line of best fit, and alternatively a technique designed to approximate the asymptotic scaling described in Sect. 8. Our code is publicly available at https://github.com/CSU-PHdimension/PHdimension.

## 5.1 Estimates of Persistent Homology Fractal Dimensions

We display several experimental results, for shapes of both integral and non-integral fractal dimension. In Fig. 3, we show the log-log plots of $L^i(X_n)$ as a function of $n$, where $X_n$ is sampled uniformly at random from a disk, a square, and an equilateral triangle, each of unit area in the plane $\mathbb{R}^2$. Each of these spaces constitutes a manifold of dimension two, and we thus expect these shapes to have persistent homology fractal dimension $d = 2$ as well. Experimentally, this appears to be the case, both for homological dimensions $i = 0$ and $i = 1$. Indeed, our asymptotically estimated slopes lie in the range 0.49–0.54, which is fairly close to the expected slope of $\frac{d-1}{d} = \frac{1}{2}$.

In Fig. 4 we perform a similar experiment for the cube in $\mathbb{R}^3$ of unit volume. We expect the cube to have persistent homology fractal dimension $d = 3$, corresponding to a slope in the log-log plot of $\frac{d-1}{d} = \frac{2}{3}$. This appears to be the case for homological dimension $i = 0$, where the slope is approximately 0.65. However, for $i = 1$ and $i = 2$, our estimated slope is far from $\frac{2}{3}$, perhaps because our computational limits do not allow us to take $n$, the number of randomly chosen points, to be sufficiently large.

In Fig. 5 we use log-log plots to estimate some persistent homology fractal dimensions of the Cantor set cross the interval (expected dimension $d = 1 + \log_3(2)$), of the Sierpiński triangle (expected dimension $d = \log_2(3)$), of Cantor dust in $\mathbb{R}^2$ (expected dimension $d = \log_3(4)$), and of Cantor dust in $\mathbb{R}^3$ (expected dimension $d = \log_3(8)$). As noted in Sect. 3, various notions of fractal dimension tend to agree for well-behaved fractals. Thus, in each case above, we provide the Hausdorff dimension $d$ in order to define an expected persistent homology fractal dimension. The Hausdorff dimension is well-known for the Sierpiński triangle, Cantor dust in $\mathbb{R}^2$, and Cantor dust in $\mathbb{R}^3$. The Hausdorff dimension for the Cantor set cross the interval can be shown to be $1 + \log_3(2)$, which follows from [30, Theorem 9.3] or [48, Theorem III]). In Sect. 5.2 we define these fractal shapes in detail, and we also explain our computational technique for sampling points from them at random.

Summarizing the experimental results for self-similar fractals, we find reasonably good estimates of fractal dimension for homological dimension $i = 0$. More

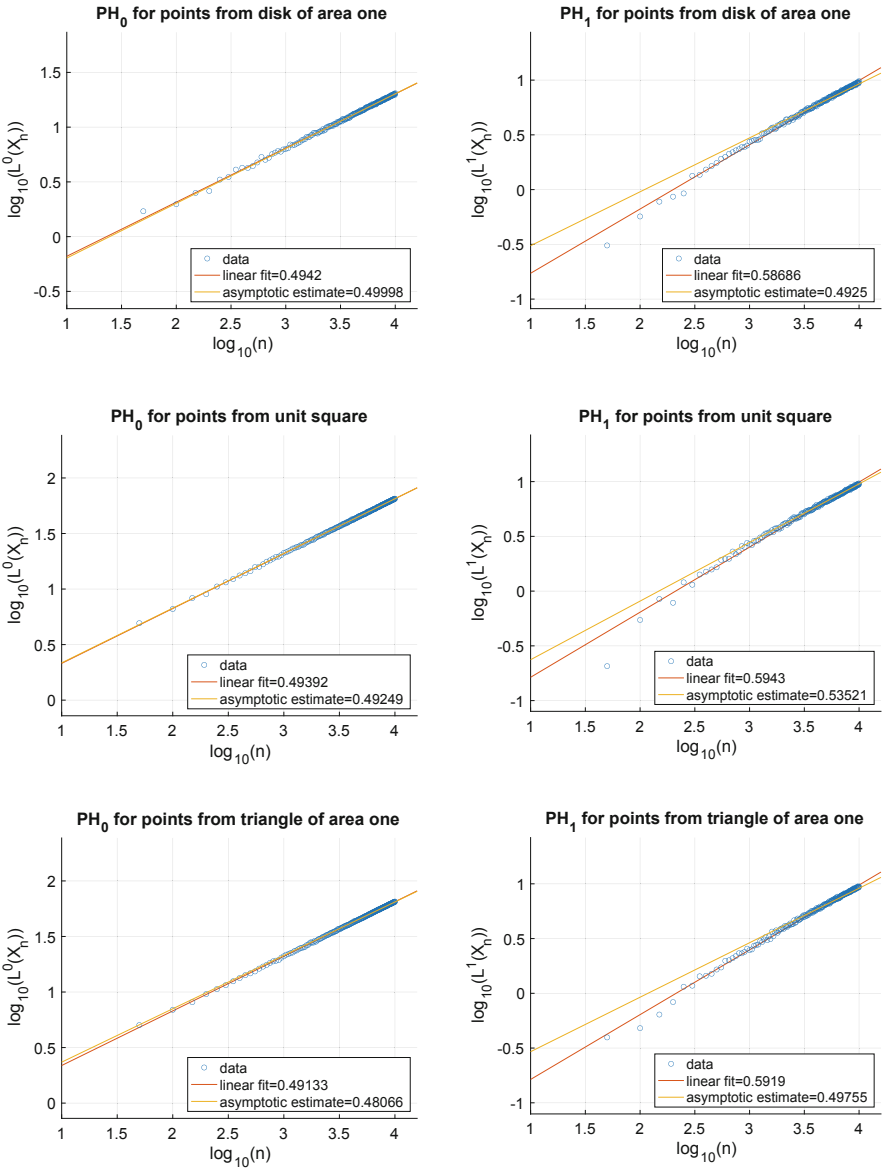


**Fig. 3** Log scale plots and slope estimates of the number $n$ of sampled points versus $L^0(X_n)$ (left) or $L^1(X_n)$ (right). Subsets $X_n$ are drawn uniformly at random from (top) the unit disc in $\mathbb{R}^2$, (middle) the unit square, and (bottom) the unit triangle. All cases have slope estimates close to $1/2$, which is consistent with the expected dimension. The asymptotic scaling estimates of the slope are computed as described in Sect. 8

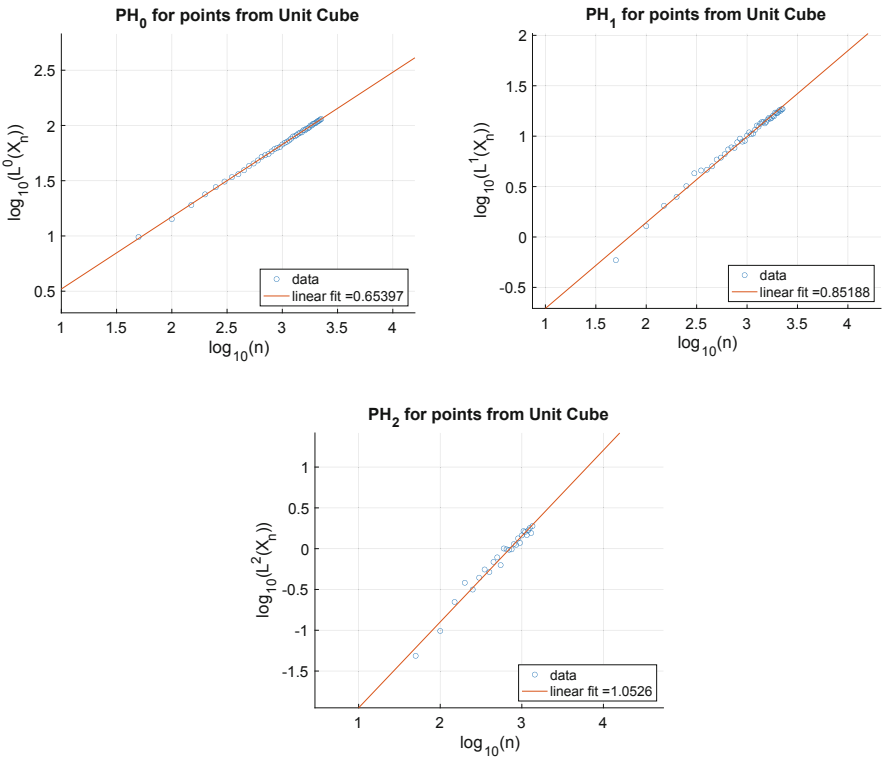


**Fig. 4** Log scale plots of the number $n$ of sampled points from the cube versus $L^0(X_n)$ (left), $L^1(X_n)$ (right), and $L^2(X_n)$ (bottom). The dimension estimate from zero-dimensional persistent homology is reasonably good, while the one- and two-dimensional cases are less accurate, likely due to computational limitations

specifically, for the Cantor set cross the interval, we expect $\frac{d-1}{d} \approx 0.3869$, and we find slope estimates from a linear fit of all data and an asymptotic fit to be 0.3799 and 0.36488, respectively. In the case of the Sierpiński triangle, the estimate is quite good: we expect $\frac{d-1}{d} \approx 0.3691$, and the slope estimates from both a linear fit and an asymptotic fit are approximately 0.37. Similarly, the estimates for Cantor dust in $\mathbb{R}^2$ and $\mathbb{R}^3$ are close to the expected values: (1) For Cantor dust in $\mathbb{R}^2$, we expect $\frac{d-1}{d} \approx 0.2075$ and estimate $\frac{d-1}{d} \approx 0.25$. (2) For Cantor dust in $\mathbb{R}^3$, we expect $\frac{d-1}{d} \approx 0.4717$ and estimate $\frac{d-1}{d} \approx 0.49$. For $i > 0$ many of these estimates of the persistent homology fractal dimension are not close to the expected (Hausdorff) dimensions, perhaps because the number of points $n$ is not large enough. The experiments in $\mathbb{R}^2$ are related to [61, Corollary 1], although our experiments are with the Vietoris–Rips complex instead of the Čech complex.

It is worth commenting on the Cantor set, which is a self-similar fractal in $\mathbb{R}$. Even though the Hausdorff dimension of the Cantor set is $\log_3(2)$, it is not hard to

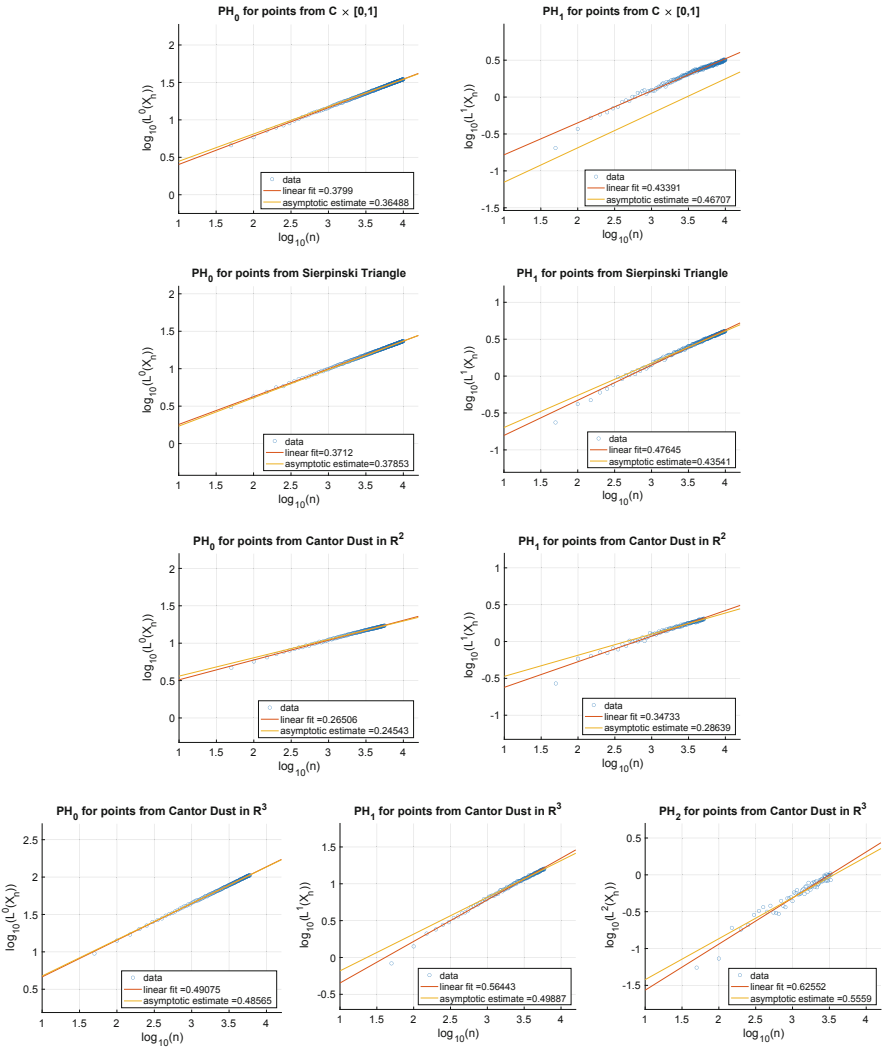


**Fig. 5** (Top) Cantor set cross the unit interval for $i = 0, 1$. (Second row) Sierpiński triangle in $\mathbb{R}^2$ for $i = 0, 1$. (Third row) Cantor dust in $\mathbb{R}^2$ for $i = 0, 1$. (Bottom) Cantor dust in $\mathbb{R}^3$ for $i = 0, 1, 2$. In each case, the zero-dimensional estimate is close to the expected dimension. The higher-dimensional estimates are not as accurate; we speculate that this is due to computational limitations

see that the zero-dimensional persistent homology fractal dimension of the Cantor set is 1. This is because as $n \to \infty$ a random sample of points from the Cantor set will contain points in $\mathbb{R}$ arbitrarily close to 0 and to 1, and hence $L_0(X_n) \to 1$ as $n \to \infty$. This is not surprising—we do not necessarily expect to be able to detect a fractional dimension less than one by using minimal spanning trees (which are one-