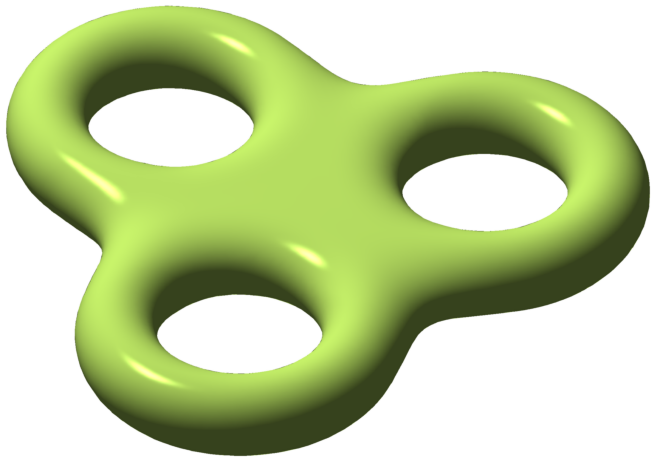
Learning Algebraic Varieties from Samples

Sara Kališnik

November 23, 2018



With Paul Breiding, Bernd Sturmfels and Madeleine Weinstein

Short overview:

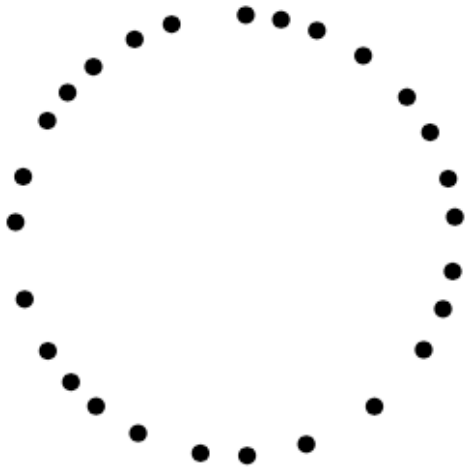
1. Introduction to Varieties (Basic Definitions, Examples, Applications)
2. Extracting Information from Samples: Dimension Estimates
3. Extracting Information from Samples: Persistent Homology
4. Extracting Information from Samples: Computing Polynomials, using Algebraic Geometry Software

Extracting Information from Samples

The Data

We are given a finite sample of points $\Omega = \{u^{(1)}, u^{(2)}, \ldots, u^{(m)}\}$ in $\mathbb{R}^n$ or $\mathbb{RP}^{n-1}$. These are sampled from an unknown variety.

Goal: Learn as much information about V as possible.



$\longrightarrow \quad (x-3)^2 + (x-5)^2 - 100$

Dimension, equations, degree, homology.

Extracting Information from Samples

Our Problem Illustrated

Input: A sample Ω of forty points in $\mathbb{R}^6$:

$(0, -2, 6, 0, -1, 12)$	$(-4, 5, -15, -12, -5, 15)$	$(-4, 2, -3, 2, 6, -1)$	$(0, 0, -1, -6, 0, 4)$
$(12, 3, -8, 8, -12, 2)$	$(20, 24, -30, -25, 24, -30)$	$(9, 3, 5, 3, 15, 1)$	$(12, 9, -25, 20, -15, 15)$
$(0, -10, -12, 0, 8, 15)$	$(15, -6, -4, 5, -12, -2)$	$(3, 2, 6, 6, 3, 4)$	$(12, -8, 9, 9, 12, -6)$
$(2, -10, 15, -5, -6, 25)$	$(5, -5, 0, -3, 0, 3)$	$(-12, 18, 6, -8, 9, 12)$	$(12, 10, -12, -18, 8, -15)$
$(1, 0, -4, -2, 2, 0)$	$(4, -5, 0, 0, -3, 0)$	$(12, -2, 1, 6, 2, -1)$	$(-5, 0, -2, 5, 2, 0)$
$(3, -2, -8, -6, 4, 4)$	$(-3, -1, -9, -9, -3, -3)$	$(0, 1, -2, 0, 1, -2)$	$(5, 6, 8, 10, 4, 12)$
$(2, 0, -1, -1, 2, 0)$	$(12, -9, -1, 4, -3, -3)$	$(5, -6, 16, -20, -4, 24)$	$(0, 0, 1, -3, 0, 1)$
$(15, -10, -12, 12, -15, -8)$	$(15, -5, 6, 6, 15, -2)$	$(-2, 1, 6, -12, 1, 6)$	$(3, 2, 0, 0, -2, 0)$
$(24, -20, -6, -18, 8, 15)$	$(-3, 3, -1, -3, -1, 3)$	$(-10, 0, 6, -12, 5, 0)$	$(2, -2, 10, 5, 4, -5)$
$(4, -6, 1, -2, -2, 3)$	$(3, -5, -6, 3, -6, -5)$	$(0, 0, -2, 3, 0, 1)$	$(-6, -4, -30, 15, 12, 10)$

Task: Learn the variety V.

Estimating the Dimension

How to use the existing literature on intrinsic dimension?

The dimensionality of a dataset is the minimum number of **free variables** needed to represent the data without information loss. In more general terms, a dataset is said to have **intrinsic dimensionality(ID)** equal to M if its elements lie entirely within an M-dimensional subspace of $\mathbb{R}^d$ (where $M < d$).

Data dimensionality estimation methods: a survey by Francesco Camastra

Maximum likelihood estimation of intrinsic dimension by E. Levina and P. Bickel

Angles and intrinsic dimension by M. Díaz, A. Quiroz and M. Velasco

Estimating the Dimension

How to use the existing literature on intrinsic dimension?

The use of more dimensions than strictly necessary leads to several problems.

- **space needed to store the data.** As the amount of available information increases, the compression for storage purposes becomes even more important.
- **slower computation time**
  The speed of algorithms using the data depends on the dimension of the vectors, so a reduction of the dimension can result in reduced computation time.
- **curse of dimensionality**
  It can be hard to make reliable classifiers when the dimensionality of input data is high.

Estimating the Dimension

How to use the existing literature on intrinsic dimension?

The known algorithms for computing intrinsic dimension of Ω can be grouped into two distinct categories: **local methods** and **global methods**.

Local methods estimates use the information contained in sample neighborhoods, whereas global approaches make use of the whole dataset. **Projection techniques** and **fractal-based methods** represent two big families of global methods.

Estimating the Dimension: Projection Techniques

Projection techniques search for the best subspace to project the data by minimizing the projection error. One such example is PCA.

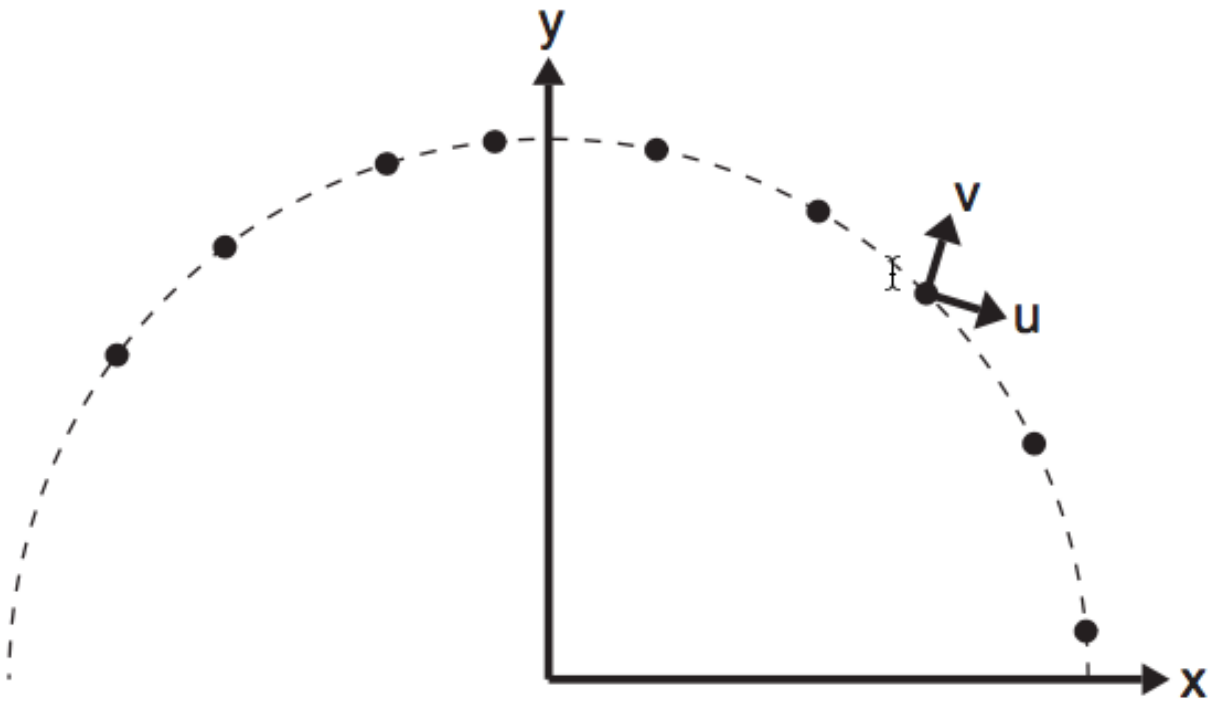
Principal Component Analysis (PCA)

Assuming that V is a linear subspace of $\mathbb{R}^n$, we perform the following steps for the input Ω.

1. We record the mean $\overline{u} := \frac{1}{m}\sum_{i=1}^{m} u^{(i)}$.
2. Let M be the $m \times n$-matrix with rows $u^{(i)} - \overline{u}$.
3. We compute $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}$, the singular values of M.
4. The PCA dimension is the number of σ_i above a certain threshold.

Each of the eigenvectors is called a **principal component**, which is where the name of the method comes from.

Estimating the Dimension: Projection Techniques



Let Ω be the data set above formed by points lying on the upper semicirconference of equation $x^2 + y^2 = 1$. The ID of is 1. Nevertheless, PCA yields two non-null eigenvalues. The principal components are indicated by $\mathbf{u}$ and $\mathbf{v}$.

Estimating the Dimension: Projection Techniques

The idea behind this NPCA is that the manifold $V \backslash \mathrm{Sing}(V)$ is approximately linear locally.

Nonlinear Principal Component Analysis (NPCA)

We partition the sample Ω into l clusters $\Omega_1^\epsilon, \ldots, \Omega_l^\epsilon \subset \Omega$ depending on ϵ. For each cluster Ω_i^ϵ we apply the usual PCA and obtain the estimate $\dim_{\mathrm{pca}}(\Omega_i^\epsilon)$. We take the average of these local dimensions, weighted by the size of each cluster. The result is the nonlinear PCA dimension

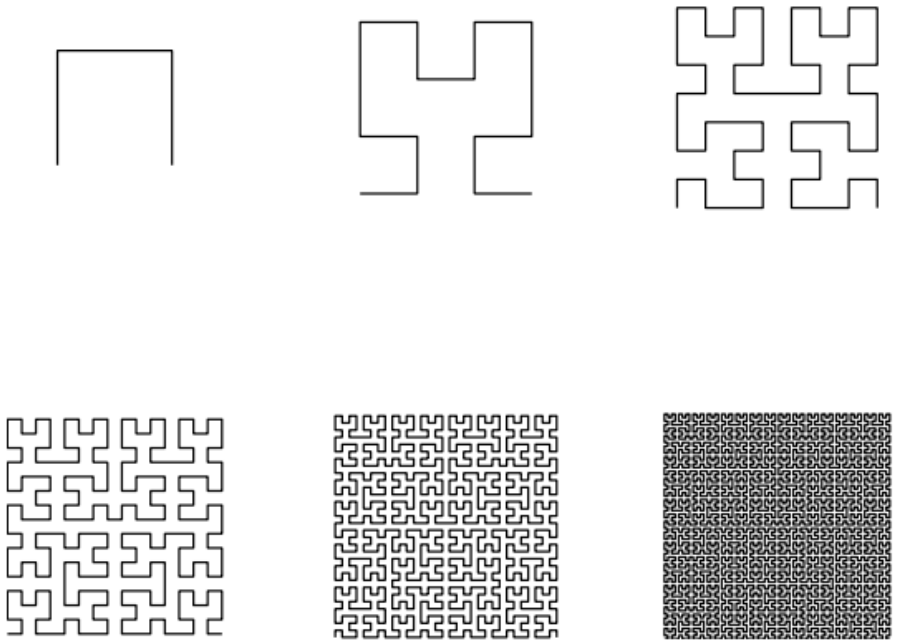
$$\dim_{\mathrm{npca}}(\Omega, \epsilon) \;:=\; \frac{1}{\sum_{i=1}^{l} |\Omega_i^\epsilon|} \sum_{i=1}^{l} |\Omega_i^\epsilon| \cdot \dim_{\mathrm{pca}}(\Omega_i^\epsilon).$$

Estimating the Dimension: Fractal-Based Methods

Motivation

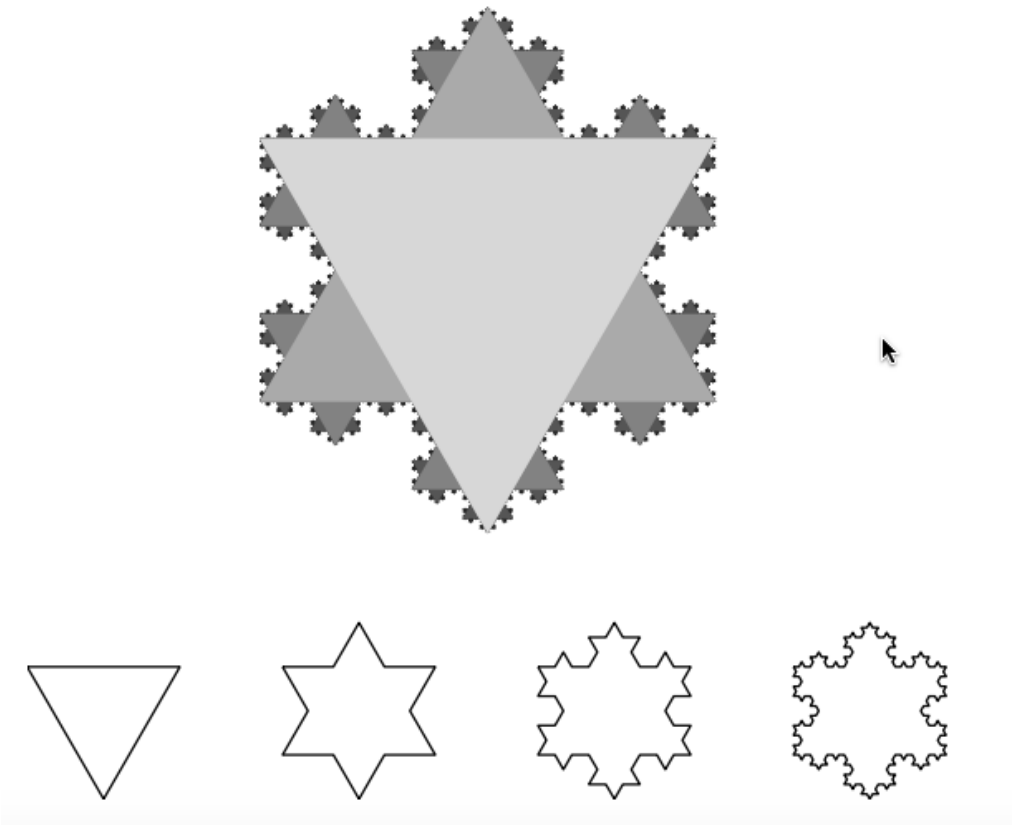
The notion of fractal dimension originates in the study of dynamical systems. The attracting sets of simple dynamical systems is often a submanifold, with an obvious dimension, but in non-linear and chaotic dynamical systems the attracting set may not be a manifold. The Cantor set, defined by removing 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.

Estimating the Dimension: Fractal-Based Methods



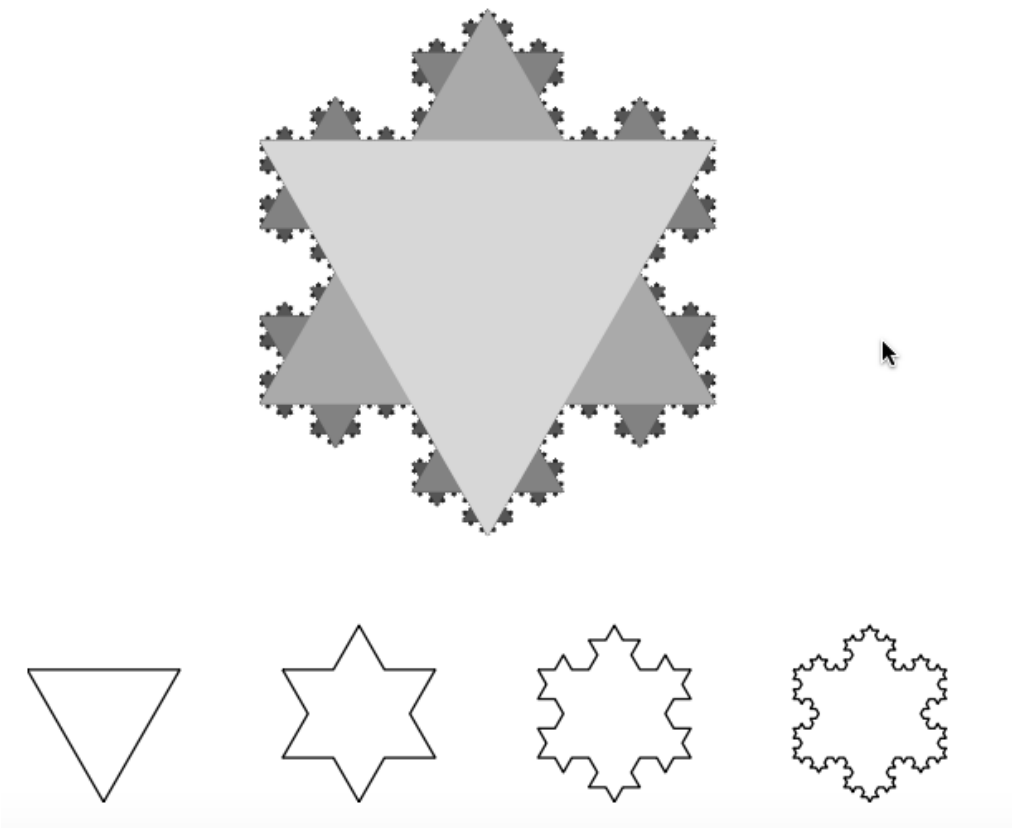
A space-filling curve. This curve, invented by Hilbert in 1891, is a one-dimensional object that evolves iteratively and progressively fills a square— a two-dimensional object! —. The first six iteration steps that are displayed show how the curve is successively refined, folded on itself in a similar way as a cabbage leaf.

Estimating the Dimension: Fractal-Based Methods



Koch's island (or snowflake). This classical fractal object was first described by Helge von Koch in 1904.

Estimating the Dimension: Fractal-Based Methods



It is built by starting with an equilateral triangle, removing the inner third of each side, replacing it with two edges of a three-times-smaller equilateral triangle, and then repeating the process indefinitely. The fractal dimension of the Koch's island is $\log_3(4)$.

Estimating the Dimension: Fractal-Based Methods

The primary definition for sets is given by the Hausdorff dimension:

Hausdorff Dimension

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

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

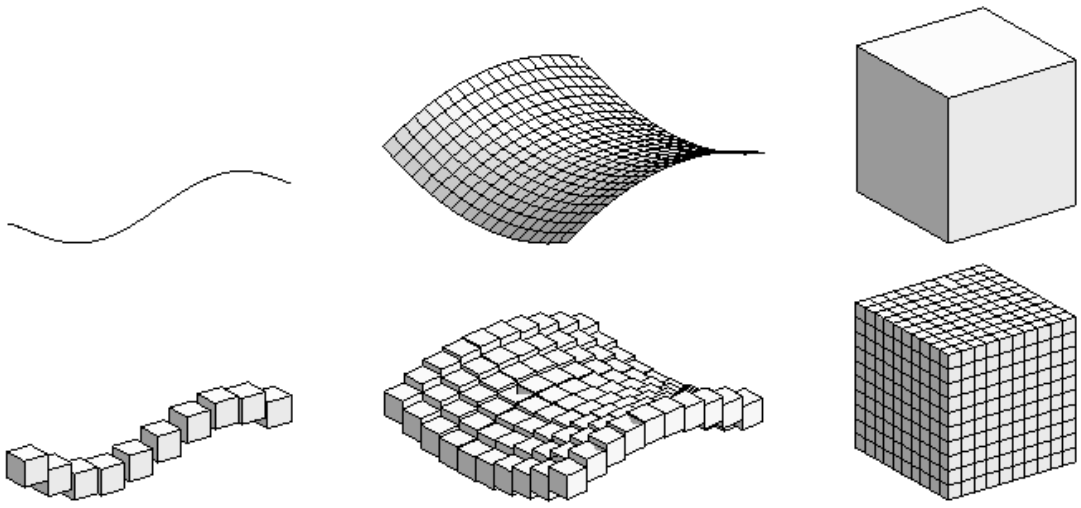
where the inner infimum is over all coverings of S by balls B_j of diameter at most δ. The Hausdorff dimension of S is

$$\dim_H(S) = \inf_d \{d \,|\, H_d(S) = 0\}$$

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

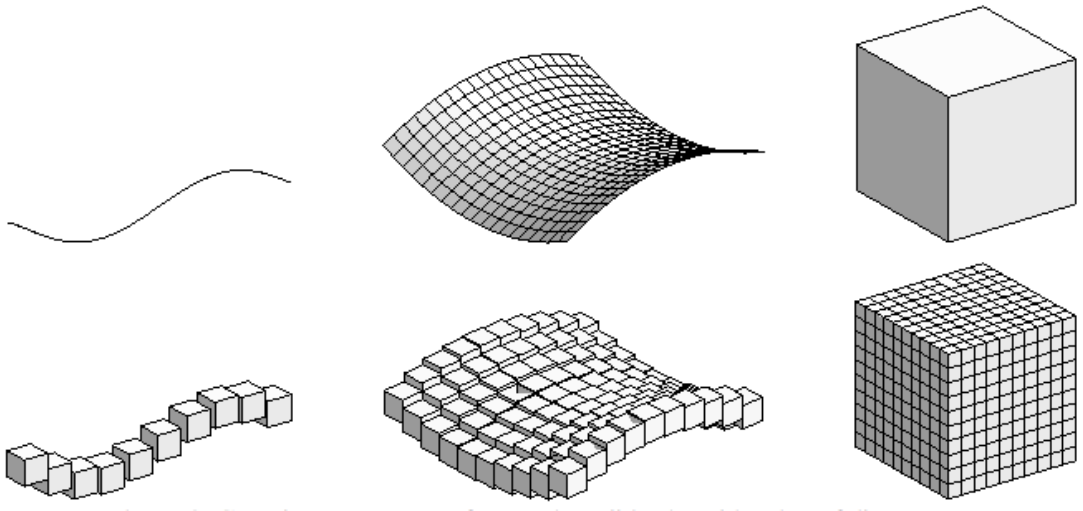
Estimating the Dimension: Fractal-Based Methods

Since the Hausdorff dimension is not easy to evaluate, in practical application it is replaced by an upper bound that differs only in pathological examples.



If we try to cover the unit square with little squares of side length ϵ, how many will we need? Obviously, the answer is $1/\epsilon^2$. How about to cover a segment of length 1? Here we need only $1/\epsilon$ little squares.

Estimating the Dimension: Fractal-Based Methods



If we think of the square and segment as sitting in space and try to cover them with little cubes ϵ on a side, we get the same answer. And if we use the little cubes to cover a $1 \times 1 \times 1$ cube, how many will we need? Exactly $1/\epsilon^3$. The exponent here is the same as the dimension of the thing we are trying to cover.

Estimating the Dimension: Fractal-Based Methods

For any $\epsilon > 0$, let $N_\epsilon(S)$ be the minimum number of n-dimensional cubes of side-length ϵ needed to cover S.
If there is a number d so that

$$N_\epsilon(S) \sim 1/\epsilon^d \qquad \text{as} \qquad \epsilon \to 0,$$

we say that the box-counting dimension of S is d.
Note that the box-counting dimension is d if and only if there is some positive constant k so that

$$\lim_{\epsilon \to 0} \frac{N_\epsilon(S)}{1/\epsilon^d} = k.$$

Since both sides of the equation above are positive, it will still hold if we take the logarithm of both sides to obtain

$$\lim_{\epsilon \to 0} (\ln N_\epsilon(S) + d \ln \epsilon) = \ln k.$$

Estimating the Dimension: Fractal-Based Methods

Solving for d gives

$$d = \lim_{\epsilon \to 0} \frac{\ln k - \ln N_\epsilon(S)}{\ln \epsilon} = -\lim_{\epsilon \to 0} \frac{\ln N_\epsilon(S)}{\ln \epsilon}.$$

Note that the $\ln k$ term drops out, because it is constant while the denominator becomes infinite as $\epsilon \to 0$. Also, since $0 < \epsilon < 1$, $\ln \epsilon$ is negative, so d is positive as we would expect.

Estimating the Dimension: Fractal-Based Methods

Box Counting Dimension

In $\mathbb{R}^n$ we choose as a box the parallelopiped with lower vertex $u^- = \min(u^{(1)}, \ldots, u^{(m)})$ and upper vertex $u^+ = \max(u^{(1)}, \ldots, u^{(m)})$, where "min" and "max" are coordinatewise minimum and maximum.

For $j = 1, \ldots, n$, the interval $[u_j^-, u_j^+]$ is divided into equally sized intervals of length ϵ (if needed, we extend the interval on the right). We determine the number $N_\epsilon(\Omega)$ of boxes that contain a point in Ω. Then the box counting dimension estimate is

$$\dim_{\text{box}}(\Omega, \epsilon) \;:=\; -\frac{\ln N_\epsilon(\Omega)}{\ln \epsilon}.$$

The projective version of this estimates involves taking the Fubini-Study distance when splitting intervals into smaller intervals.

Estimating the Dimension: Fractal-Based Methods

Box Counting Dimension and Persistent Homology

Let $MST(X)$ denote the minimal spanning tree of a finite point set X in a metric space and let

$$E_d^0(X) = \frac{1}{2} \sum_{e \in MST(X)} \|e\|^d$$

where the sum is taken over all edges e in the tree $MST(X)$, and $\|e\|$ denotes the length of the edge.
Define $\dim_{MST}(X)$ to be

$$\dim_{MST}(X) = \inf\{d : E_d^0(\{x_j\}) < C \,\forall \,\text{finite subsets } \{x_j\} \text{ of } X\}.$$

Then

$$\dim_{MST}(X) = \overline{\dim_{box}}(X).$$

The minimal spanning tree and the upper box dimension by G. Kozma, Z. Lotker and G. Stupp

Estimating the Dimension: Fractal-Based Methods

Box Counting Dimension and Persistent Homology

For finite metric spaces X there is a bijection between the edges of the Euclidean minimal spanning tree of X and the intervals in the canonical decomposition of $\mathrm{PH}_0(X)$ (with Čech complexes), where the length of an interval in $\mathrm{PH}_0(X)$ is half the length of the corresponding edge. In this spirit, B. Scheweinhart defined

$$E_d^i(X) = \sum_{(x,y)\in PH_i(X)} (y-x)^d$$

where the sum is taken over all bounded $PH_i(X)$ intervals, and

$$\dim_{PH}^i(X) = \inf\{d \,:\, E_d^i(\{x_j\}) < C \,\forall \,\text{finite subsets } \{x_j\} \text{ of } X\}.$$

Persistent Homology and the upper box dimension by B. Scheweinhart

Estimating the Dimension: Fractal-Based Methods

A good substitute for the box-counting dimension can be the **correlation dimension**. The correlation dimension gives a lower bound on the Hausdorff dimension of a measure.

Correlation Dimension

Let $\Omega = \{u^{(1)}, u^{(2)}, \ldots, u^{(m)}\}$ be a set of points in $\mathbb{R}^n$ of cardinality m. The correlation integral $C(\epsilon)$ is defined as

$$C_2(\epsilon) = \lim_{m \to \infty} \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} I(\|(u^{(j)} - u^{(i)}\| \leq \epsilon),$$

where I is an indicator function.

Estimating the Dimension: Fractal-Based Methods

Correlation Dimension

When looking at the data set on the scale of a single point, $C_2(\epsilon)$ is the number of neighboring points lying closer than a certain threshold ϵ. This number grows as a length for a 1D object, as a surface for a 2D object, as a volume for a 3D object, and so forth. So we expect $C(\epsilon)$ to be approximately ϵ^d. Just like before this suggests using $\frac{\log(C(\epsilon))}{\log(\epsilon)}$ as a dimension estimate.

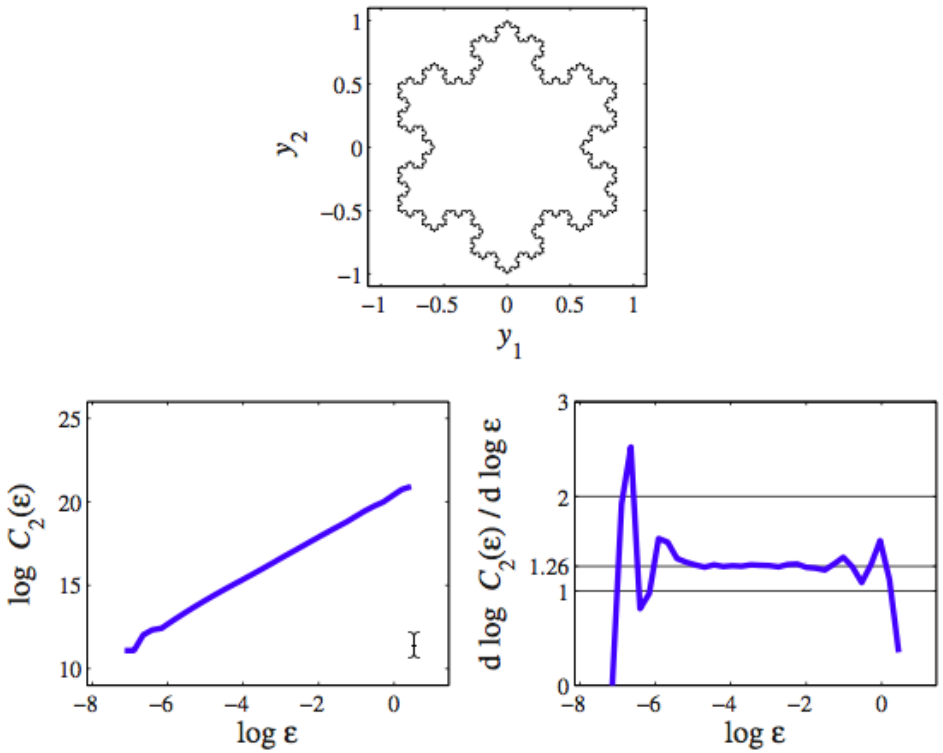
A more practical estimate is obtained from $C(\epsilon)$ by selecting some small $h > 0$ and putting

$$\dim_{\text{cor}}(\Omega, \epsilon) \ := \ \left| \frac{\log C(\epsilon) - \log C(\epsilon + h)}{\log(\epsilon) - \log(\epsilon + h)} \right|.$$

In practice, we compute the dimension estimates for a finite subset of parameters $\epsilon_1, \ldots, \epsilon_k$ and put $h = \min_{i \neq j} |\epsilon_i - \epsilon_j|$.

Estimating the Dimension: Fractal-Based Methods

Correlation Dimension



Usually, the best estimate for the dimension is obtained in the largest region where the slope of $C_2(\epsilon)$ computed from the sample is almost constant in the log-log plot. This region is often called a "plateau".

Estimating the Dimension: Fractal-Based Methods

Persistent Homology Curve Dimension Estimate

First we partition Ω into l clusters $\Omega_1^\epsilon, \ldots, \Omega_l^\epsilon$ using single linkage clustering with ϵ. On each subsample Ω_i we construct a minimal spanning tree. Suppose that the cluster Ω_i has m_i points. Let $f_i(j)$ be the length of the j-th longest edge in a minimal spanning tree for Ω_i. For each Ω_i we compute

$$\dim_{\mathrm{PHcurve}}(\Omega_i, \epsilon) = \left| \frac{\log(m_i)}{\log(\frac{1}{m_i - 1} \sum_{j=1}^{m_i - 1} f_i(j))} \right|.$$

The persistent homology curve dimension estimate $\dim_{\mathrm{PHCurve}}(\Omega, \epsilon)$ is the average of the local dimensions, weighted by the size of each cluster:

$$\dim_{\mathrm{PHcurve}}(\Omega, \epsilon) := \frac{1}{\sum_{i=1}^{l} |\Omega_i^\epsilon|} \sum_{i=1}^{m} |\Omega_i| \dim_{\mathrm{PHcurve}}(\Omega_i, \epsilon).$$

Estimating the Dimension

Maximum Likelihood Estimation of Intrinsic Dimension

Let k be the number of samples $u^{(j)}$ in Ω that are within distance ϵ to $u^{(\star)}$. We write $T_i(u^{(\star)})$ for the distance from $u^{(\star)}$ to its i-th nearest neighbor in Ω. Note that $T_k(u^{(\star)}) \leq \epsilon < T_{k+1}(u^{(\star)})$. The *Levina-Bickel formula* around the point $u^{(\star)}$ is

$$\dim_{\mathrm{MLE}}(\Omega, \epsilon, u^{(\star)}) \;:=\; \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{\epsilon}{T_i(u^{(\star)})} \right)^{-1} .$$

This expression is derived from the hypothesis that $k = k(\epsilon)$ obeys a Poisson process on the ϵ-neighborhood $\{u \in \Omega : \mathrm{dist}_{\mathbb{R}^n}(u, u^{(\star)}) \leq \epsilon\}$, in which u is uniformly distributed. The formula is obtained by solving the likelihood equations for this Poisson process.

Estimating the Dimension

Maximum Likelihood Estimation of Intrinsic Dimension

It is not clear how to choose $u^{(\star)}$ from the given Ω. We chose the following method. We fix the sample neighborhood $\Omega_i^\epsilon := \{u \in \Omega : \mathrm{dist}_{\mathbb{R}^n}(u, u^{(i)}) \leq \epsilon\}$. For each i we evaluate the formula for Ω_i^ϵ with distinguished point $u^{(i)}$. With this, the *MLE dimension estimate* is

$$\dim_{\mathrm{MLE}}(\Omega, \epsilon) \;:=\; \frac{1}{\sum_{i=1}^m |\Omega_i^\epsilon|} \sum_{i=1}^m |\Omega_i^\epsilon| \cdot \dim_{\mathrm{MLE}}(\Omega_i^\epsilon, \epsilon, u^{(i)}).$$

Estimating the Dimension

ANOVA Dimension

For data $u^{(1)}, \ldots, u^{(m)}$ sampled from a manifold $M \subset \mathbb{R}^n$ and a point $u^{(\star)} \in M$ Diaz, Quiroz and Velasco consider a statistic which estimates the variance of the angle between pairs of vectors $u^{(i)} - u^{(\star)}$ and $u^{(j)} - u^{(\star)}$, for data points $u^{(i)}$, $u^{(j)}$, near $u^{(\star)}$ and evaluate this statistic as a tool for estimation of the intrinsic dimension of M at $u^{(\star)}$.
For uniform data on S^{d-1}, the expected angle between two random vectors is always $\frac{\pi}{2}$ (regardless of d), but the variance β_d of this angle decreases rapidly with d.

$$\beta_{2s-1} = \frac{\pi^2}{4} - 2\sum_{j=0}^{s} \frac{1}{(2j+1)^2} \quad \text{and} \quad \beta_{2s} = \frac{\pi^2}{12} - 2\sum_{j=0}^{s} \frac{1}{(2j)^2}.$$

for $s \in \mathbb{N}$.

Estimating the Dimension

ANOVA Dimension

We again fix $\epsilon > 0$, and we relabel so that $u^{(1)}, \ldots, u^{(k)}$ are the points in Ω with distance at most ϵ from $u^{(\star)}$. Let $\theta_{ij} \in [0, \pi]$ denote the angle between $u^{(i)} - u^{(\star)}$ and $u^{(j)} - u^{(\star)}$. Then, they define the **angle-variance** of the θ_{ij} as

$$S \;=\; \frac{1}{\binom{k}{2}} \sum_{1 \le i < j \le k} \left(\theta_{ij} - \frac{\pi}{2}\right)^2.$$

For small ϵ and Ω sampled from a d-dimensional manifold, the angles θ_{ij} are approximately Θ_d-distributed. Hence, S is expected to be close to $\beta_{\dim M}$. The **ANOVA dimension estimate** of Ω is the index d such that β_d is closest to S:

$$\dim_{\mathrm{ANOVA}}(\Omega, \epsilon, u^{(\star)}) \;:=\; \operatorname{argmin}_d |\beta_d - S|.$$

Estimating the Dimension

How to use the existing literature on intrinsic dimension?

Key point: Our sample size $m = |\Omega|$ is fixed and relatively small.

There are various estimators $\dim^*(\Omega, \epsilon)$. These depend on a parameter $\epsilon > 0$ and they produce positive real numbers.

Key point: ϵ does not tend to 0. This would be meaningless.

We tackle this by instead considering dimension diagrams. The dimension diagram of the sample Ω is the graph of the function $(0, 1) \to \mathbb{R}_{\geq 0}$, $\epsilon \mapsto \dim(\Omega, \epsilon)$, where $\dim(\Omega, \epsilon)$ is a dimension estimate.

Examples and Software

The implementations are available in the `Julia` package `LearningAlgebraicVarieties`.

We offer a step-by-step tutorial. To install our software, start a `Julia` session and type

```
Pkg.clone("https://github.com/PBrdng/LearningAlgebraicVarieties.git")
```

After the installation, the next command is

```
using LearningAlgebraicVarieties
```

This command loads all the functions into the current session. Our package accepts a dataset Ω as a matrix whose *columns* are the data points $u^{(1)}, u^{(2)}, \ldots, u^{(m)}$ in $\mathbb{R}^n$.

To use the numerical algebraic geometry software **Bertini**, we must first download it from `https://bertini.nd.edu/download.html`. The `Julia` wrapper for **Bertini** is installed by

```
Pkg.clone("https://github.com/PBrdng/Bertini.jl.git")
```

Sample Datasets

Our software includes samples from the following varieties:

- $\mathrm{SO}(3)$;
- the projective variety of 2×3-matrices of rank 1;
- the conformation space of the cyclo-octane molecule;
- 3×4 rank two matrices.

We provide the samples used in the subsequent experiments in the JLD[2] data format. After having installed the JLD package in Julia (`Pkg.add("JLD")`), load the datasets by typing

```
import JLD: load
s = string(Pkg.dir("LearningAlgebraicVarieties"),"/datasets.jld")
datasets = load(s)
```

Analyzing a Sample from SO(3)

To produce our sample from $\mathrm{SO}(3)$, we sample a $\mathbb{R}^{3\times 3}$ matrix from a standard Gaussian and then take a Q of the QR-decomposition.

```
data = datasets["SO(3)"]
```

Now the current session should contain a variable `data` that is a 9×887 matrix. We produce the dimension diagrams by typing

```
DimensionDiagrams(data, false, methods=[:CorrSum,:
    PHCurve])
```

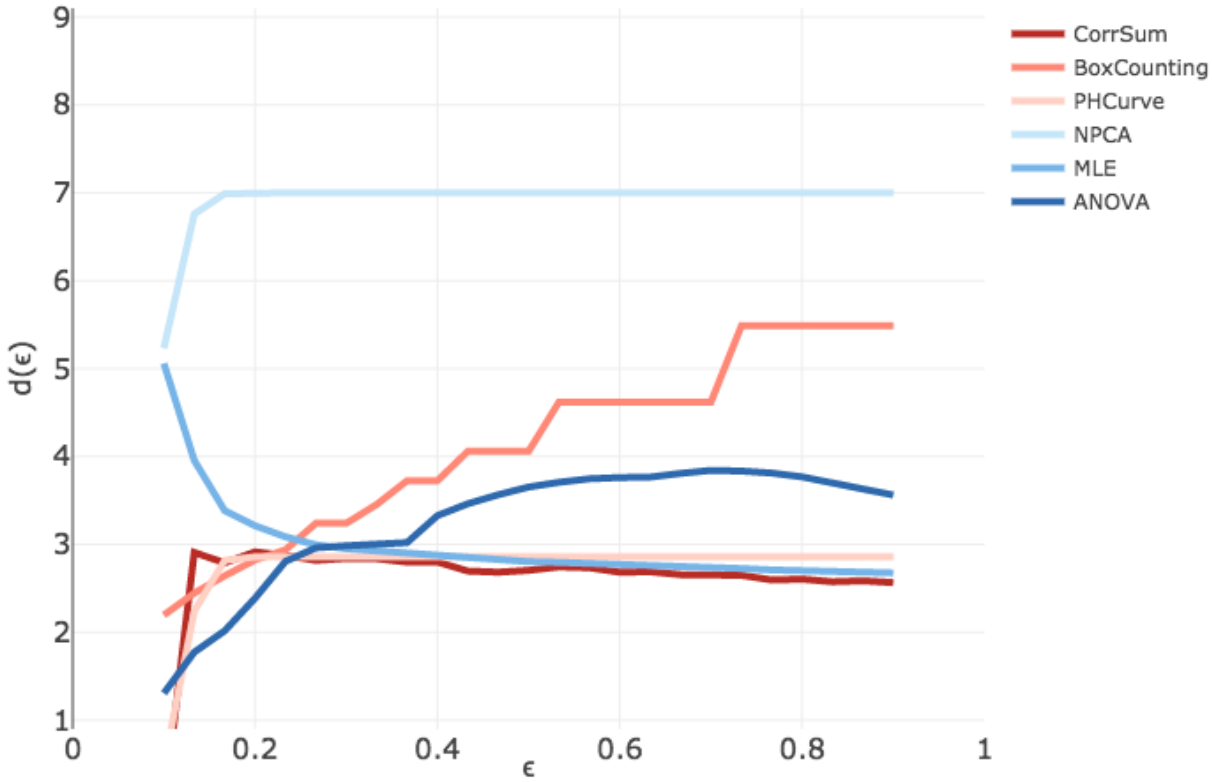
In this command, `data` is our dataset, the Boolean value is **`true`** if we suspect that our variety is projective and **`false`** otherwise, and `methods` is any of the dimension estimates `:CorrSum`, `:BoxCounting` `:PHCurve`, `:NPCA`, `:MLE`, and `:ANOVA`. We can leave this unspecified and type

```
DimensionDiagrams(data, false)
```

This command plots all six dimension diagrams.

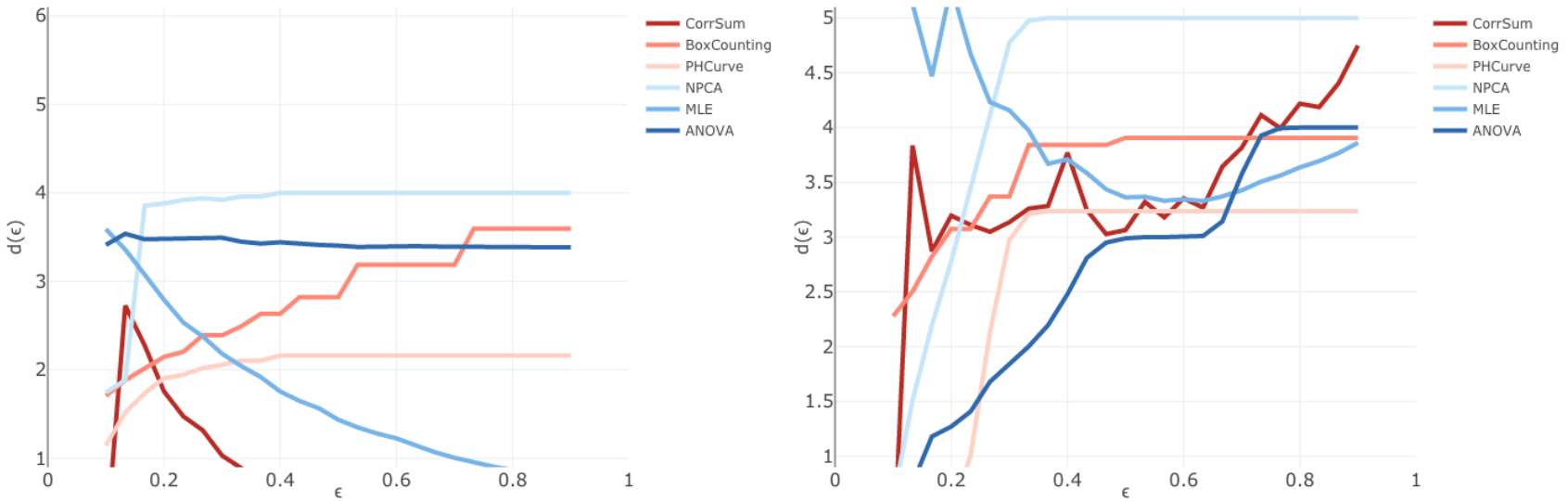
Analyzing a Sample from SO(3)

NPCA often overestimates the dimension as is the case with our sample from $\mathrm{SO}(3)$. Three estimates are very close to 3 so we guess that the dimension is 3, which is in fact the true dimension. If the dimension diagrams were more spread out, then we can still use them to get a range of values that are candidates for the dimension.



Analyzing a Sample from the Segre Variety

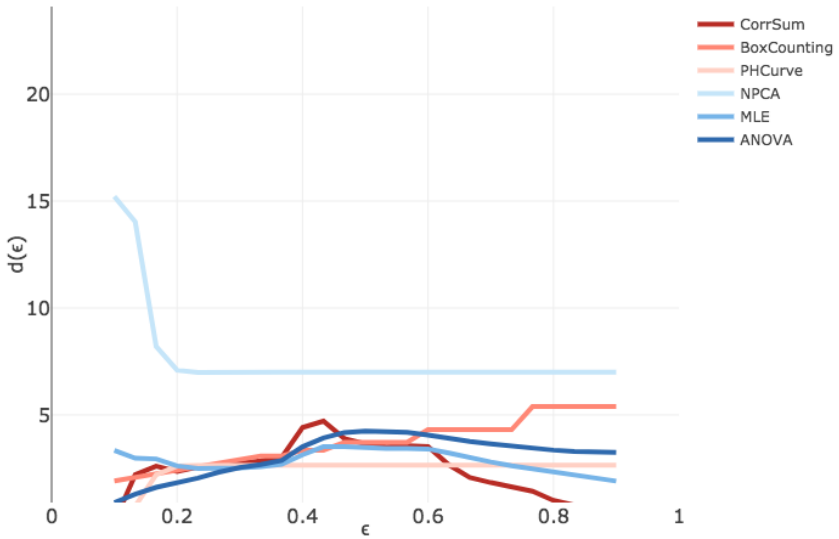
We produce a sample for the Segre variety $V = \mathbb{RP} \times \mathbb{RP}^2$ by independently sampling two standard Gaussian matrices of format 2×1 and 1×3 and multiplying them.



Dimension diagrams for 200 points on the variety of 2×3 matrices of rank 1. The left picture shows dimension diagrams for the estimates in $\mathbb{R}^6$. The right picture shows those for projective space $\mathbb{RP}^5$. The true dimension is 3.

Analyzing a Sample from Cyclo-octane

We use the same sample of 6040 points that was analyzed by Martin et al and randomly select a.



Dimension diagrams for 420 randomly selected points from the cyclo-octane sample. The true dimension is 2.