A tutorial for topological data analysis 1

Otto van Koert

Seoul National University, South Korea

- Topological data analysis (TDA) is a relatively new field in mathematics with the promise of many potential new applications.
- Philosophy summarized by Carlsson as: data has shape, and shape has meaning.
- Today: Mapper and persistent homology, and some sample applications.

Main philosophy

- Geometry and topology are subfields of mathematics that are concerned with shape
- Geometry measures quantitative properties: distance, volume, etc
- Topology is concerned with more qualitative properties, and algebraic topology transforms these into computable algebraic language.
- Applications of these techniques should be backed up with the appropriate statistics. Today we ignore these statistical issues.

Shape of data

Data has sometimes a rather simple shape as indicated below. Regression is an excellent technique in such cases.

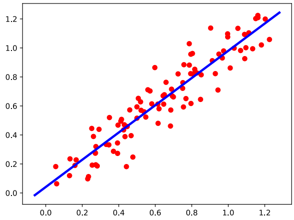


Figure: Noisy linear data

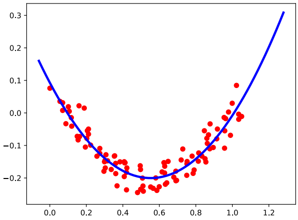


Figure: Noisy quadratic data

Other shapes

More complicated shapes appear, too.

Figure: the (transformed) amount versus time of the transaction

This shape appears in transaction data: the interpretation is obvious: each of the seven clusters corresponds to a week day: the smaller clusters are Saturday and Sunday.

Clustering

Statistics is well-equipped to deal with the previous example: the subfield of clustering deals with this particular problem. For example, most of the following methods will work

- hierarchical clustering (single-linkage, complete linkage)
- *k*-means
- distribution-based methods such as expectation-maximization (EM)
- density-based methods (DB-scan)

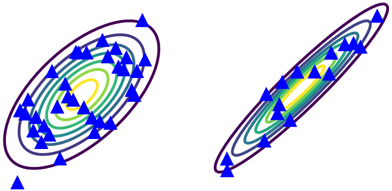


Figure: Clusters found with EM

A tutorial for topological data analysis 1

Otto van Koert

Some applications

More complicated shapes

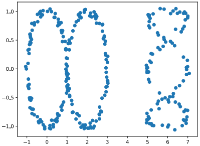
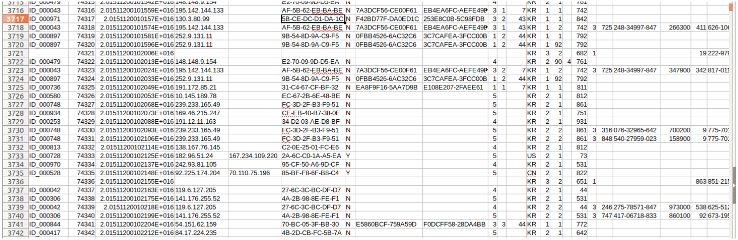


Figure: Several of the above clustering algorithms will fail here



3716	ID_000043	74316	2.01511200101559E+016	195.142.144.133		AF-5B-62-EB-BA-BE	N	7A3DCF56-CE00F61	EB4EA6FC-AEFE49	3	1	7	KR	1	1		742						
3717	ID_000971	74317	2.0151120010157E+016	130.3.80.99		5B-CE-DC-D1-DA-1C	N	F42BD77F-DA0ED1C	253E8C0B-5C98FDB	3	2	43	KR	1	1		842						
3718	ID_000043	74318	2.01511200101574E+016	195.142.144.133		AF-5B-62-EB-BA-BE	N	7A3DCF56-CE00F61	EB4EA6FC-AEFE49	3	1	43	KR	1	2		742	3	725	248-34997-847	266300	411	626-10
3719	ID_000897	74319	2.01511200101581E+016	252.9.131.11		9B-54-8D-9A-C9-F5	N	0FBB4526-6AC32C6	3C7CAFEA-3FCC00B	1	2	44	KR	1	1		792						
3720	ID_000897	74320	2.01511200101596E+016	252.9.131.11		9B-54-8D-9A-C9-F5	N	0FBB4526-6AC32C6	3C7CAFEA-3FCC00B	1	2	44	KR	1	92		792						
3721		74321	2.01511200102006E+016										KR	3	2		682	1				19	222-97
3722	ID_000479	74322	2.01511200102013E+016	148.148.9.154		E2-70-09-9D-D5-EA	N			4			KR	2	90	4	761						
3723	ID_000043	74323	2.01511200102024E+016	195.142.144.133		AF-5B-62-EB-BA-BE	N	7A3DCF56-CE00F61	EB4EA6FC-AEFE49	3	2	7	KR	1	2		742	3	725	248-34997-847	347900	342	817-01
3724	ID_000897	74324	2.01511200102033E+016	252.9.131.11		9B-54-8D-9A-C9-F5	N	0FBB4526-6AC32C6	3C7CAFEA-3FCC00B	1	2	44	KR	1	92		792						
3725	ID_000736	74325	2.01511200102049E+016	191.172.85.21		31-C4-67-CF-BF-32	N	EA8F9F16-5AA7D9B	E108E207-2FAEE61	1	1	7	KR	1	1		811						
3726	ID_000580	74326	2.01511200102053E+016	10.145.189.78		EC-67-2B-6E-48-BE	N			5			KR	2	1		812						
3727	ID_000748	74327	2.01511200102068E+016	239.233.165.49		FC-3D-2F-B3-F9-51	N			5			KR	2	1		861						
3728	ID_000934	74328	2.01511200102073E+016	169.46.215.247		CE-EB-40-B7-38-0F	N			5			KR	2	1		751						
3729	ID_000253	74329	2.01511200102088E+016	191.12.11.163		34-D2-03-AE-D8-BF	N			4			KR	2	1		931						
3730	ID_000748	74330	2.01511200102093E+016	239.233.165.49		FC-3D-2F-B3-F9-51	N			5			KR	2	2		861	3	316	076-32965-642	700200	9	775-70
3731	ID_000748	74331	2.01511200102106E+016	239.233.165.49		FC-3D-2F-B3-F9-51	N			5			KR	2	2		861	3	848	540-27959-023	158900	9	775-70
3732	ID_000813	74332	2.01511200102114E+016	138.167.76.145		C2-0E-25-01-FC-E6	N			4			KR	2	1		812						
3733	ID_000728	74333	2.01511200102125E+016	182.96.51.24	167.234.109.220	2A-6C-C0-1A-A5-EA	Y			5			US	2	1		73						
3734	ID_000970	74334	2.01511200102137E+016	242.93.81.105		95-CF-50-A6-9D-CF	N			4			KR	2	1		531						
3735	ID_000528	74335	2.01511200102148E+016	92.225.174.204	70.110.75.196	85-BF-F8-6F-B8-C4	Y			5			CN	2	1		822						
3736		74336	2.01511200102155E+016										KR	3	2		651	1				863	851-21
3737	ID_000042	74337	2.01511200102163E+016	119.6.127.205		27-6C-3C-BC-DF-D7	N			4			KR	2	1		44						
3738	ID_000306	74338	2.01511200102175E+016	141.176.255.52		4A-2B-98-8E-FE-F1	N			5			KR	2	1		531						
3739	ID_000042	74339	2.0151120010218E+016	119.6.127.205		27-6C-3C-BC-DF-D7	N			4			KR	2	2		44	3	246	275-78571-847	973000	538	625-51
3740	ID_000306	74340	2.01511200102199E+016	141.176.255.52		4A-2B-98-8E-FE-F1	N			5			KR	2	2		531	3	747	417-06718-833	860100	92	673-19
3741	ID_000844	74341	2.01511200102204E+016	54.151.62.159		70-BC-05-3F-BB-30	N	E5860BCF-759A59D	F0DCFF58-28DA4BB	3	3	44	KR	1	1		772						
3742	ID_000417	74342	2.01511200102212E+016	84.17.224.235		4B-2D-CB-FC-5B-7A	N			5			KR	2	1		642						

Figure: (fake) transaction data: many categorical variables

The previous example highlights some of the problems with modern data.

- a large number of data points
- many different coordinates (or features) with no direct meaning.
- related: distance does not have a direct meaning.

The last two points mean that direct geometric methods may give weak or unstable results. TDA complements these methods, because

- topology is by definition independent of the metric and
- shapes have different behavior at different scales: this will be captured with the notion of *persistence*, which relies on *functoriality*.

Topology

We will need a couple of notions of topology, such as connectedness (previous example). We will also need a suitable class of topological spaces with a good decomposition

1. CW-complexes: preferred in algebraic topology. Examples include general graphs
2. Simplicial complexes: decomposes a topological space into many, simple pieces. This class also includes many graphs, but not all.

Simplicial complexes seem to be most useful for computational work; we will use that class after a quick review.

Simplex as a generalized triangle

Definition

A geometric k-**dimensional simplex** is the topological space

$$\Delta_k := \{(x_0, \ldots, x_k) \in \mathbb{R}^{k+1} \mid \sum_{i=0}^{k} x_i = 1,\ x_i \geq 0\}$$

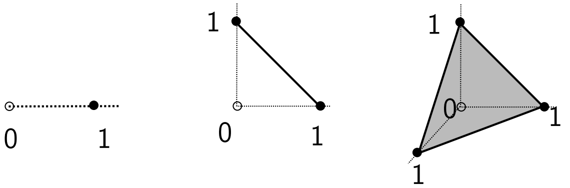
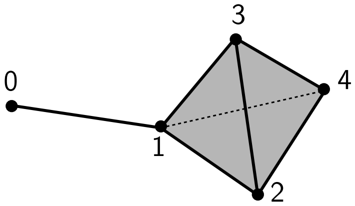


Figure: A 0-simplex, a 1-simplex and a 2-simplex: a point, a line segment and a triangle

Simplicial complexes and abstract simplicial complexes

- Roughly put, a simplicial complex is a topological space constructed with simplices using combinatorial "gluing recipe". See the figure below.
- An abstract simplicial complex is this combinatorial recipe; it is ideal for computers, but actually does not need any geometry.



Simplicial complexes: Lego-like construction descriptions of spaces

Definition (Slightly confusing)

An **abstract simplicial complex** is a (finite) collection of (finite) sets X such that if $x \in X$, and $y \subset x$, then $y \in X$.

Example

$X = \{\{0\}, \{1\}, \{2\}, \{0, 1\}\}$
We see that the definition holds.

For practical purposes, it is useful to think of $X = X_0 \cup X_1 \cup \ldots \cup X_n$, where

- X_k consists of sets with $k + 1$ elements: X_k parametrizes the k-simplices.

In the example we have

$$X_0 = \{\{0\}, \{1\}, \{2\}\}, \quad X_1 = \{\{0, 1\}\}$$

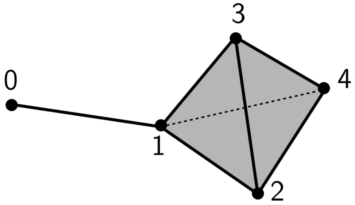
Here is a more complicated example of an abstract simplicial complex.

$$X_0 = \{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}\}$$

$$X_1 = \{\{0,1\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\}$$

$$X_2 = \{\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}\}$$

We can construct a topological space out of this by replacing each k-simplex by a geometric k-simplex.



Remark

For later purposes (orientations) we will always order vertices in a simplex by increasing index.

Geometric realization:

Given an abstract simplicial complex $X = \{X_n\}_{n=0}^N$, we will define the **geometric realization** of X. Intuitively, this is just the shape built from the Lego description.
Here is a simple way to make it explicit:

1. order X_0
2. choose N sufficiently large and an embedding $i : X_0 \to \mathbb{R}^N$ such that $\{i(x)\}_{x \in X_0}$ are linearly independent (can be weakened)
3. for each $\sigma \in X_k$ we get a map $f_\sigma : \Delta_k \to \mathbb{R}^N$ by linear combination. Namely if $x \in \Delta_k$, then $x = \sum_{j=0}^k t_j e_j$, where e_j is the standard basis of $\mathbb{R}^{k+1}$.
   Put $f_\sigma(x) = \sum_j t_j i(\sigma[j])$, where $\sigma[j]$ is the j-th point of σ.
4. the geometric realization is $|X| = \cup_{\sigma \in X} f_\sigma(\Delta_{|\sigma|})$

Remark

In step 2 we asked for linear independence to prevent unwanted intersections between the images of different simplices. In most cases, linear independence is much stronger than necessary.

Remark

We will never really need geometric realization. The abstract simplicial complex suffices for all computational work.

Remark

Ordering the vertices, i.e. X_0*, is the standard way to deal with orientations (which we only discuss implicitly).*

Definition

A (finite) simplicial complex is a topological space that is homeomorphic to the geometric realization of a (finite) abstract simplicial complex.

Simplicial complexes and compression

Many interesting (but not all) topological spaces admit a simplicial structure. If this is the case, we can see such a simplicial structure as a compressed representation:

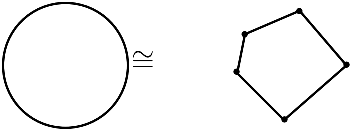


Figure: Circle as a simplicial complex

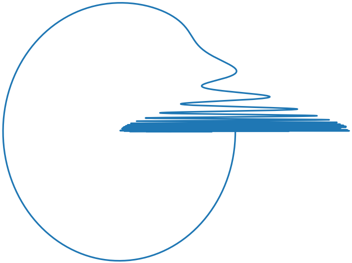


Figure: The quasi-circle has no simplicial structure

Nerve of a covering

Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a covering of a topological space X.

Definition

The **nerve of the covering** $\mathcal{U}$ is the simplicial complex $N(\mathcal{U})$, whose simplices are defined as follows:

- vertices are $N(\mathcal{U})^0 = \{u_i\}_{i \in I}$, so $U_i \in \mathcal{U}$ gives one vertex.
- $[u_0, \ldots, u_k]$ forms a k-simplex when $\cap_{j=0}^{k} U_j \neq \emptyset$.

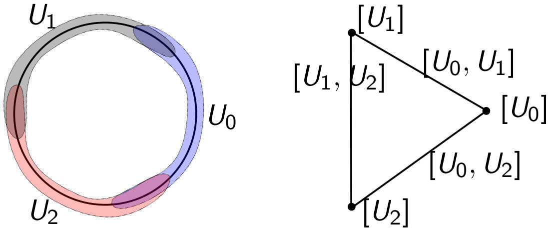


Figure: A covering of the circle and its nerve

Nerve theorem

Theorem (Nerve theorem)

Suppose that X is a paracompact topological space, and assume that $\mathcal{U} = \{U_i\}_i$ is a good cover, meaning that any finite intersection

$$U_{i_0} \cap \ldots \cap U_{i_k}$$

is either empty or contractible. Then the nerve $\mathcal{N}(\mathcal{U})$ is homotopy equivalent to X.

Roughly speaking, this theorem says that the nerve of a good cover can be deformed to the original space.

Remark

The mathematical-philosophical background (more later) is the following. Most invariants from algebraic topology depend only on the homotopy type of a space. So replacing a space by a homotopy equivalent one that is easier, helps computations.

A brief excursion to the Reeb graph

Suppose that X is a topological space, and $f : X \to \mathbb{R}$ a continuous function. Define an equivalence relation on X by

$x \sim_f x'$ if and only if x, x' are in the same component of $f^{-1}(y)$.

Define the **Reeb graph** of (X, f) as

$$R_f(X) := X / \sim_f$$

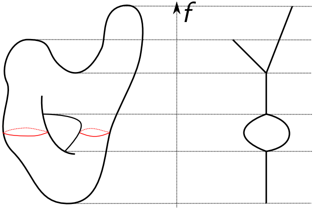


Figure: Reeb graph of a function

Pullback covers

We have the following relation between the Reeb graph and the nerve construction. Fix a continuous function $f : X \to \mathbb{R}$. Call f a filter function or lens.

- Cover $f(X)$ by intervals I_i.
- The sets $U_i = f^{-1}(I_i)$ form a cover (almost never good)
- To improve our "chances", we decmpose U_i into its connected components $U_i = \cup_j C_{i,j}$. If the space X is reasonable (locally connected), then $C_{i,j}$ are open.

This results in an open cover $X = \cup_{i,j} C_{i,j}$.

- we can now look at the nerve construction for this cover
- and to the Reeb graph of f

If the cover by intervals I_i is sufficiently fine, we get the "same" result.

Reeb graph and mapper

Remark

The Reeb graph captures a rough version of the topology: a lot of information is lost, but some is retained in a graph, which is computationally easier to deal with.

We cannot directly apply the Reeb graph to point cloud data; with only finitely many points in the cloud, most level sets are empty. Hence replace a level-set by the preimage of an interval.

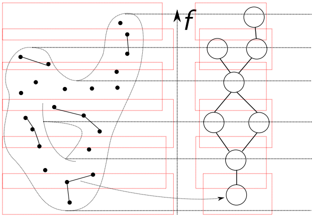
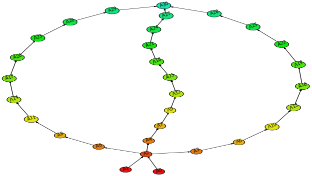


Figure: An analog of the Reeb graph

..Mapper

1. fix a data set and a function (filter function) from the data set to $\mathbb{R}$.
2. cover $\mathbb{R}$ with intervals that overlap on a smaller interval (gain)
3. put points whose filter value lies in interval I_b in bin b.
4. cluster points in each bin b: these clusters are the vertices of the mapper
5. connect the mapper vertices if the corresponding clusters share a point.



Multi-dimensional mapper

For the higher-dimensional version of mapper, the nerve construction is our guide. Suppose we want to construct a k-dimensional mapper for a pointcloud X. Then

1. choose filter functions $f_i : X \to \mathbb{R}$ for $i = 1, \ldots k$. Collect them in $F : X \to \mathbb{R}^k$
2. cover $F(X)$ with overlapping boxes U_i
3. to obtain the analog of the pullback cover, cluster data points in each U_i. Call the clusters $C_{i,j}$
4. The clusters form the vertices of mapper
5. There is an edge if two clusters $C_{i,j}$ and $C_{i',j'}$ overlap (i.e. share a data point) This can be generalized to several clusters to get higher-dimensional simplices.

The main purposes of mapper are:

- visualization
- exploratory data analysis

It is flexible and can deal with large and diverse data sets: mixed numerical and categorical data can be dealt with. One only needs a dissimilarity measure.

Figure: A model of a human

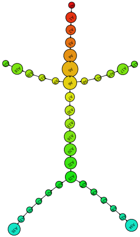


Figure: Mapper of the first PCA-function for a human model; PCA stands for principal component analysis

Choices

There are lots of choices in mapper. This gives it a lot of flexibility, but this comes at a price, namely complexity.

- We need to choose filter functions of interest
- The number of bins, and the epsilon parameters need to be chosen.

Here it should be pointed out that statistics gives us some guidance.

- more bins speeds up the mapper algorithm
- but more bins results in fewer points per bin, which makes the statistical data (mean, etc per bin) less reliable.

Some applications

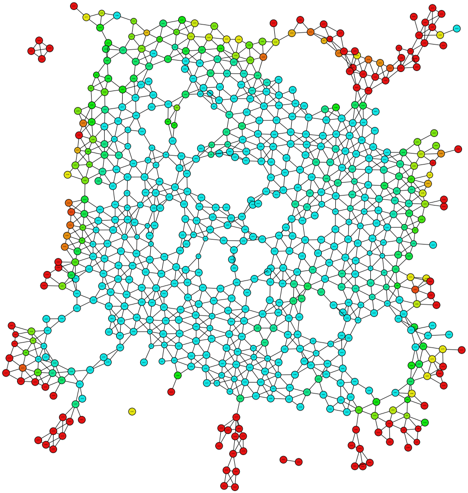
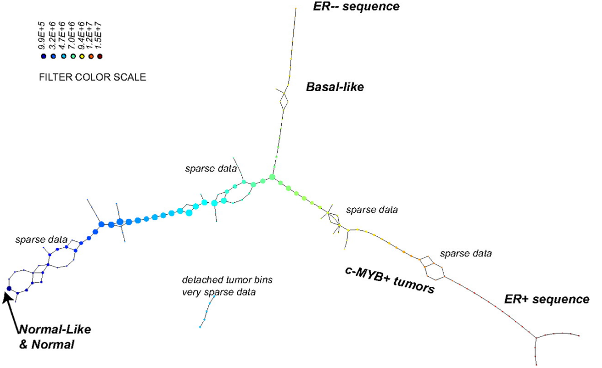
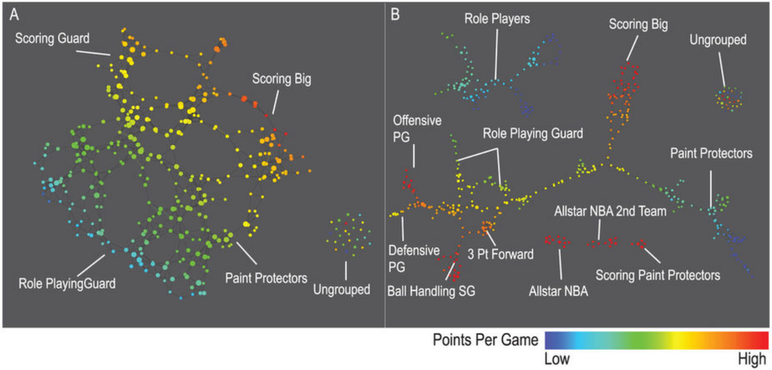


Figure: Mapper of medical data with two filter functions, and colored by a third filter function (disease: yes or no)

The following picture is taken from
Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival
by Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson, PNAS April 26, 2011. 108 (17) 7265-7270;



The following picture is taken from *Extracting insights from the shape of complex data using topology*
by P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson and G. Carlsson, Nature, Scientific Reports volume 3, Article number: 1236



Thank you
감사 합니다.