Data and homotopy types

Rick Jardine

University of Western Ontario

May 13, 2019

Basic setup

$X \subset \mathbb{R}^n$ a finite subset (data set, or data cloud).

X is finite: $\{0 = s_0, s_1, s_2, \ldots, s_p\}$ = distances between points of X, with $s_i < s_j$ if $i < j$ (**phase changes**).

X is listed on a computer, so $X \cong \{0, \ldots, N\}$ (listing is an **orientation**).

Vietoris-Rips complex $V_s(X)$: k-simplex $\sigma = \{x_0, \ldots, x_k\}$ (ordered, or not), with $d(x_i, x_j) \leq s$.

$V_s(X) \subset V_t(X)$ if $s < t$.

$V_0(X) = X$. If $s_p \leq R$, then $V_R(X) = \Delta^N =: \Delta^X$, $|X| = N + 1$.

$$X = V_0(X) \subset \cdots \subset V_s(X) \subset \cdots \subset V_t(X) \subset \cdots \subset V_R(X) = \Delta^X$$

$s \mapsto V_s(X)$ defines a functor $V_*(X) : [0, R] \to \mathbf{Set}$.

Persistent homology

Write $H_p(Y) = H_p(Y, k)$, $k =$ a field, $p \geq 1$.

We have a filtration

$$k(X) = k(V_0(X)) \subset \cdots \subset k(V_s(X)) \subset \cdots \subset k(V_R(X)) = k(\Delta^X)$$

of the chain complex $k(\Delta^X)$, and induced maps

$$H_p(V_0(X)) \to \cdots \to H_p(V_s(X)) \to \cdots \to H_p(V_R(X)) = H_p(\Delta^X).$$

We care only about the phase changes and the maps

$$0 = H_p(V_0(X)) \to H_p(V_{s_1}(X)) \to \cdots \to H_p(V_{s_p}(X)) = H_p(\Delta^X) = 0.$$

This is a fin. dim. module over $k[t]$. The structure theorem for finitely generated modules over a p.i.d. breaks this "persistence module" up into a direct sum of principal modules, which give the bar codes ....

Clusters

Apply the path component functor to the maps

$$X = V_0(X) \subset \cdots \subset V_s(X) \subset \cdots \subset V_R(X) = \Delta^X$$

to produce functions

$$X = \pi_0 V_0(X) \to \cdots \to \pi_0 V_s(X) \to \cdots \to \pi_0 V_R(X) = \pi_0(\Delta^X) = *$$

All spaces $V_s(X)$ have the same vertices, namely X, and $\pi_0 V_s(X)$ is a partition of X, a **clustering** of the data set X.

If $s \leq t$, then $\pi_0 V_s(X) \to \pi_0 V_t(X)$ is surjective, and the partition given by $\pi_0 V_s(X)$ is a refinement of that given by $\pi_0 V_t(X)$.

Hierarchy graph: $\Gamma(X) := \Gamma(\pi_0 V_*(X))$ with vertices $(s, [x])$, $[x] \in \pi_0 V_s(X)$, and morphisms $(s, [x]) \to (t, [x])$ for $s \leq t$ in $\mathbb{R}$.

$\Gamma(X)$ is a **tree** (dendogram), a **hierarchical clustering** for X.

Issues

1) Specifying a distance s between stars in the picture X determines a Vietoris-Rips complex $V_s(X)$ and path components (clusters) $\pi_0 V_s(X)$.

If s is small (but not too small), the clusters consist of a big central blob and outlying groups.

2) Varying s a little does not change the path component picture. Components that remain intact through variations of s are **stable components** (also **stable clusters**, **layers**) — these are persistence objects.

3) X is the result of setting an exposure time. A longer exposure produces Y with more points, and $X \subset Y$: $\Gamma(X) \to \Gamma(Y)$.

Stability question: How well do the two trees approximate each other?

Practical issues

Given $X \subset \mathbb{R}^n$, the $V_s(X)$ are found as follows:

1) Compute all distances $d(x, y)$ for $x, y \in X$.
2) Consider all finite subsets $\sigma = \{x_0, x_1, \ldots, x_p\}$ and compute $s = \max\{d(x_i, x_j)\}$. Then $\sigma \in V_s(X)$.

This algorithm has exponential complexity.

A properly equipped PC (16GB RAM, 1TB SSD) can only handle 1,000 data points at once, and then only for low dimensional simplices of the $V_s(X)$. Handling 10,000 points requires more sophisticated hardware.

Need local to global methods to handle larger data sets.

mapper only gives approximate calculations.

Sheaves [3]

What about this functor $s \mapsto V_s(X)$? What are its global homotopy theoretic properties?

Restrict to functors $[0, R] \to s\mathbf{Set}$, such as $s \mapsto V_s(X)$, where $R \geq s_p$, so $V_0(X) = X$ and $V_R(X) = \Delta^X$.

Suppose that $Y : [0, R] \to \mathbf{Set}$.

$I \subset [0, R]$ is an interval:

A **persistent element** u on I is a string of elements $u_s \in Y_s$, $s \in I$, which is compatible in the sense that $u_s \mapsto u_t$ for $s \leq t$ in I. ie. $u \in \varprojlim_{s \in I} Y(s)$.

If $I \subset J$ are intervals in $[0, R]$ then there is a restriction map

$$\varprojlim_{t \in J} Y(t) \to \varprojlim_{s \in I} Y(s).$$

We have a presheaf $\varprojlim Y$ on $[0, R]$.

Sheaves and stalks

$Y : [0, R] \to \mathbf{Set}$:

1) $\varprojlim Y$ is a sheaf on $[0, R]$

2) $(\varprojlim Y)_t :=$ stalk of $\varprojlim Y$ at $t \in [0, R]$:

- $(\varprojlim Y)_t \cong \varinjlim_{s<t} Y(s)$ if $t \in (0, R]$
- $(\varprojlim Y)_0 = Y(0)$.

$V_*(X) : [0, R] \to s\mathbf{Set}$ determines a simp. sheaf $\varprojlim V_*(X)$.

Quillen model structure on $s\,\mathrm{Pre}([0, R])$: a map $X \to Y$ is a weak equivalence if and only if all maps $X_t \to Y_t$ in stalks are weak equivalences of simplicial sets

(stalkwise, or local weak equivalence).

This idea **fails** for data comparisons: Given $X \subset Y \subset \mathbb{R}^n$,
$\varprojlim V_*(X) \to \varprojlim V_*(Y)$ is a weak equivalence if and only if $X = Y$.

Fuzzy sets

Traditional: a fuzzy set is a function $\phi : X \to [0, 1]$

Given $\psi : Y \to [0, 1]$, a **morphism**

$$f : \phi \to \psi$$

of fuzzy sets consists of a function $f : X \to Y$ and a relation (homotopy) $\phi \leq \psi \cdot f$ of functions taking values in $[0, 1]$.

i.e. $\phi(x) \leq \psi(f(x))$ for all $x \in X$.

Revision (Barr, 1986): $[0, 1]$ is a locale (nice poset). Fuzzy sets are functions $X \to L$, where L is some locale.

Fuzz(L) is the category of fuzzy sets $X \to L$ with values in L.

Locales

A **locale** L is a poset with infinite joins (unions) and finite meets (intersections), in which finite meets distribute over all joins.

NB: L has a terminal object (empty meet), an initial object (empty join), and infinite meets.

A **morphism** $L_1 \to L_2$ of locales is a poset morphism $L_2 \to L_1$ which preserves meets and joins (hence preserves initial and terminal objects).

Note the variance ... and of course there is a category of locales.

Examples

1) $op|_X$ = open subsets of a topological space X is a locale.

2) $[0,1]$ is a locale, as is any closed interval $[a,b] \subset \mathbb{R}$ (has initial and terminal objects).

3) A finite product $L_1 \times \cdots \times L_k$ of locales L_i is a locale.

4) The opposite poset $[0,R]^{op}$ is a locale.

5) L a locale: $L_+ = \{0\} \sqcup L$ (new disjoint initial object) is a locale.

Write i for the initial object of L.

$$0 < i \text{ in } L_+.$$

Why anyone cares:

Every locale L has a Grothendieck topology (as a category). The family $b_i \leq a$ **covers** a if $\vee_i\ b_i = a$.

Fuzzy sets and sheaves

Given a fuzzy set $\phi : X \to L$, form pullbacks

$$\begin{array}{ccc} \phi^{-1}(L_{\geq a}) & \xrightarrow{\subset} & X \\ \downarrow & & \downarrow \phi \\ L_{\geq a} & \xrightarrow[\subset]{} & L \end{array}$$

where $L_{\geq a} = \{x \,|x \geq a\}$ for $a \in L$.

$$a \mapsto \phi^{-1}(L_{\geq a}) =: T(\phi)(a)$$

defines a sheaf on L_+. $T(\phi)(0) = *$. $T(\phi)$ is a sheaf of monomorphisms on L_+: $\mathbf{Mon}(L_+)$ = sheaves of monomorphisms.

2) A sheaf F of monomorphisms on L_+ has a **generic fibre** $F(i)$. Given $x \in F(i)$ there is a min $s_x \in L$ such that $x \in F(s_x)$.
$x \mapsto s_x$ defines $\phi : F(i) \to L$.

Theorem (Barr, 1986 [1]) There is an equivalence of categories

$$\mathbf{Fuzz}(L) \simeq \mathbf{Mon}(L_+).$$

Vietoris-Rips sheaves

$X \subset \mathbb{R}^n$: Let $\sigma = \{x_0, \ldots, x_k\}$ be a set of points in X. Set

$$\phi(\sigma) = \min_{i,j} \{ \ d(x_i, x_j) \ \}.$$

We have a function $\phi : \Delta^X_k \to [0, R]^{op}$, and the corresponding sheaf has $T(\phi)(s) = V_s(X)$.

$V_*(X)$ is a simplicial sheaf on the locale $[0, R]^{op}_+$, aka. a simplicial fuzzy set.

Recall: X is finite, so there is a list

$$0 = s_0 < s_1 < \cdots < s_p \leq R$$

of all distances between points of X — **phase change** numbers.

Stalks, again

Sheaves (and simplicial sheaves) F on $[0, R]_+^{op}$ have stalks:

$$F_s = \varinjlim_{t<s} F(t), \quad s > 0,$$

(relation in $[0, R]$) and $F_0 = F(R)$ is the generic fibre.

$V_*(X)_s = V_s(X)$ if $s \notin \{s_0, \ldots, s_p\}$.

We lose again: If $X \subset Y \subset \mathbb{R}^n$ then $V_*(X) \to V_*(Y)$ is a local weak equivalence of simplicial sheaves on $[0, R]^{op}$ if and only if $X = Y$.

For $s \neq 0$ sufficiently small (i.e. less than all non-zero phase change numbers for Y), there is a commutative diagram

$$\begin{array}{ccc} X & \longrightarrow & Y \\ \cong\downarrow & & \downarrow\cong \\ V_s(X) & \to & V_s(Y) \end{array}$$

General picture

Generic example: $i : X \subset Y \subset \mathbb{R}^n$ (inclusion of finite data sets).

Naive tools of simplicial sheaf homotopy theory do not work for studying induced comparison $i : V_*(X) \to V_*(Y)$, because these systems involve discrete spaces.

What is true: all induced maps $V_s(X) \to V_s(Y)$ are weak equivalences for s sufficiently large:

If s is larger than all phase change numbers for Y (hence for X), then $V_s(X) = \Delta^X$, $V_s(Y) = \Delta^Y$, and both spaces are contractible.

r-density

$X \subset Y \subset \mathbb{R}^n$. Suppose $r > 0$.

Say that X is r-**dense** in Y if for every point $y \in Y$ there is an $x \in X$ such that $d(x, y) < r$.

Define $\theta : Y \to X$ by specifying that $\theta(y)$ is a nearest neighbour of y in x.

Remarks:

- $d(y, \theta(y)) < r$ for all $y \in Y$.
- If $y \in X$, then $\theta(y) = y$.
- If $d(y_1, y_2) \leq s$, then $d(\theta(y_1), \theta(y_2)) \leq s + 2r$ (triangle inequality).
- X is r-dense in Y if r is sufficiently large.
- If X is r-dense in Y and r is sufficiently small, then $X = Y$.

Subdivisions

$\theta : Y \to X$ induces a simplicial complex map $V_s(Y) \to V_{s+2r}(X)$, or rather a functor

$$\theta : NV_s(Y) \to NV_{s+2r}(X)$$

of posets of simplices (order complexes, barycentric subdivisions)

$$\sigma = \{y_0, \ldots, y_k\} \mapsto \theta(\sigma) = \{\theta(y_0), \ldots, \theta(y_k)\}$$

There is a diagram

$$\begin{array}{ccc} NV_s(X) & \xrightarrow{\alpha} & NV_{s+2r}(X) \\ {\scriptstyle i}\downarrow & \nearrow_{\theta} & \downarrow{\scriptstyle i} \\ NV_s(Y) & \xrightarrow[\alpha]{} & NV_{s+2r}(Y) \end{array} \tag{1}$$

in which the upper triangle commutes, and the lower triangle commutes up to homotopy:

$$\{y_0, \ldots, y_k\} \subset \{y_0, \ldots, y_k, \theta(y_0), \ldots, \theta(y_k)\} \supset \{\theta(y_0), \ldots, \theta(y_k)\}$$

Persistence functors

The construction (1) induces "interleavings" of persistence modules

$$\begin{array}{ccc} H_k V_s(X) & \xrightarrow{\alpha} & H_k V_{s+2r}(X) \\ \downarrow{\scriptstyle i} & \nearrow{\scriptstyle \theta} & \downarrow{\scriptstyle i} \\ H_k V_s(Y) & \xrightarrow[\alpha]{} & H_k V_{s+2r}(Y) \end{array}$$

and of clusters

$$\begin{array}{ccc} \pi_0 V_s(X) & \xrightarrow{\alpha} & \pi_0 V_{s+2r}(X) \\ \downarrow{\scriptstyle i} & \nearrow{\scriptstyle \theta} & \downarrow{\scriptstyle i} \\ \pi_0 V_s(Y) & \xrightarrow[\alpha]{} & \pi_0 V_{s+2r}(Y) \end{array}$$

Both diagrams commute on the nose, and are natural in s.

Homotopy type stability

Given $X \subset Y \subset \mathbb{R}^n$, let $0 = s_0 < s_1 < \cdots < s_p$ be the phase change numbers for Y.

Lemma 1.

Suppose that $2r < s_{k+1} - s_k$.

Then $i : V_t(X) \to V_t(Y)$ *is a weak equivalence for* $s_k \leq t < s_{k+1}$

Proof.

$V_t(X) = V_{s_k}(X)$ (same for Y), so it suffices to show that i is a weak equivalence for $t = s_k$.

$$\begin{array}{ccc} NV_{s_k}(X) & \overset{=}{\to} & NV_{s_k+2r}(X) \\ {\scriptstyle i}\downarrow & \overset{\theta}{\nearrow} & \downarrow{\scriptstyle i} \\ NV_{s_k}(Y) & \underset{=}{\to} & NV_{s_k+2r}(Y) \end{array}$$

homotopy commutes, so $V_{s_k}(X)$ is a def. retract of $V_{s_k}(Y)$. □

Remarks

Lemma: Suppose that $2r < s_{k+1} - s_k$.

Then $i : V_t(X) \to V_t(Y)$ is a weak equivalence for $s_k \leq t < s_{k+1}$

1) If $2r < s_0$, then $X = Y$, because of the diagram

$$\begin{array}{ccc} X & \overset{\equiv}{\to} & V_{2r}(X) \\ \downarrow & \nearrow & \downarrow \\ Y & \underset{\equiv}{\to} & V_{2r}(Y) \end{array}$$

2) The placement of $2r \geq s_0$ relative to the (ordered) differences $s_{k+1} - s_k$, $k \geq 0$, determines a set $\{ s_k \}$ for which

$$V_{s_k}(X) \to V_{s_k}(Y)$$

are weak equivalences. This includes sufficiently large s_k.

Clusters revisited

Data set $X \subset \mathbb{R}^n$ defines sets of clusters $\pi_0 V_s(X)$ and the **hierarchy graph** (category, tree, dendogram) $\Gamma(X) := \Gamma \pi_0 V_*(X)$.

$\Gamma(X)$ is the graph whose objects are pairs $([x], s)$ with $[x] = [x]_s \in \pi_0 V_s(X)$, and has edges $([x]_s, s) \to ([x]_t, t)$ with $s \leq t$.

$\Gamma(X)$ has a subgraph (subcategory) $\Gamma_0(X)$ having the same objects, and with edges $([x]_s, s) \to ([x]_t, t)$ such that $s \leq t$ and $[x]_s = [x]_t$ as subsets of X.

This is the **stable component graph** ("layer graph", "stable cluster graph") for X. Path components of $\Gamma_0(X)$ are the **stable components** (layers) of X.

Branch points

$([x]_t, t)$ is a **branch point** if the preimage of $([x]_t, s)$ under the map $\pi_0 V_s(X) \to \pi_0 V_t(X)$ has more than one element for all $s < t$ (so t has to be a phase change number).

$([x]_s, s) \to ([x]_t, t)$ is in $\Gamma_0(X)$ if and only if $([x]_v, v)$ is not a branch point of $\Gamma(X)$ for $s < v \leq t$.

All layers $L \subset \Gamma_0(X)$ have the form

$$L = \{([x]_s, s) \mid u \leq s < v\}$$

$([x]_u, u)$ and $([x]_v, v)$ are branch points.

Observations

1) A layer L is determined by the branch point $([x]_u, u)$ at which it "starts".

2) Can identify layers with branch points, including all $(\{x\}, 0)$.

Comparison of hierarchies

$i : X \subset Y \subset \mathbb{R}^n$, X r-dense in Y:

1) Every $([y], s)$ lifts to $([\theta(y)], s + 2r)$.

2) $\pi_0 V_s(X) \to \pi_0 V_s(Y)$ is surjective for $s \geq 2r$.

3) Given $([x], s), ([y], s)$, if $([i(x)], t) = ([i(y)], t])$ for some $t \geq s$, then $([x], t + 2r) = ([y], t + 2r)$.

4) If $([i(x)]_s, s)$ is a branch point of $\Gamma(Y)$ and if $s > 2r$ (so that $\pi_0 V_s(X) \to \pi_0 V_s(Y)$ is surjective), then $([x]_t, t)$ is a branch point for some $s \leq t \leq s + 2r$.

Lemma 2.

If $s > 2r$, every branch point $([y], s)$ of $\Gamma(Y)$ has a branch point $([x], t)$ of $\Gamma(X)$ that is "nearby" in sense that $s \leq t < s + 2r$.

Blackboard example

Convergence

The hierarchy $\Gamma(X)$ determines an **ultrametric** on X: say that

$$d_{\Gamma(X)}(x, y) = d(x, y) = \min\{r \mid [x]_r = [y]_r\}.$$

$$d(x, z) \leq \max\{d(x, y), d(y, z)\}$$

Theorem: (Carlsson-Memoli) $d_{GH}(sl(X), sl(Y)) \leq d_{GH}(X, Y)$ for single linkage clustering sl on X, Y.

$X \subset Y \subset \mathbb{R}^n$, X r-dense in Y:

X and Y are finite metric spaces.

Hausdorff distance: $d_H(X, Y) = \max\{d(x, y) \mid x \in X, y \in Y\}$ for imbeddings $X, Y \subset Z$.

Our setting: $X \subset Y$, X r-dense in Y: $d_H(X, Y) < r$.
$d_{GH}(X, Y) \leq d_H(X, Y) < r$.

Observation: $d_{\Gamma(X)}(x, y)$ is a "convergence rate". Convergence is faster in $\Gamma(X)$ than in $sl(X)$.

Convergence comparisons

$X \subset Y \subset \mathbb{R}^n$, X r-dense in Y:

$x, y \in X$: $d_{\Gamma(X)}(x, y)$ is smallest s such that $[x]_s = [y]_s$.

$d_{\Gamma(Y)}(x, y) \leq d_{\Gamma(X)}(x, y)$, but

$$d_{\Gamma(X)}(x, y) \leq d_{\Gamma(Y)}(x, y) + 2r.$$

Extend to all vertices of $\Gamma(X)$: $d(([x]_s, s), ([y]_t, t))$ is the minimal r such that $[x]_r = [y]_r$.

1) $d(([x]_s, s), ([y]_t, t)) = u$ if both vertices are in the same layer $L = \{([x]_p, p) \mid u \leq p < v\}$.

2) $d(([x]_s, s), ([y]_t, t))$ is invariant of layer representatives in both variables. d is defined on $\pi_0\Gamma_0(X)$.

Induced function $i_* : \pi_0\Gamma_0(X) \to \pi_0\Gamma_0(Y)$: $i_*([x], s)$ (branch point) is the layer containing $([i(x)], s)$.

$$d(i_*([x], s), i_*([y], t)) \leq d(([x], s), ([y], t)).$$

References


Michael Barr.
Fuzzy set theory and topos theory.
Canad. Math. Bull., 29(4):501–508, 1986.


Andrew J. Blumberg and Michael Lesnick.
Universality of the homotopy interleaving distance.
CoRR, abs/1705.01690, 2017.


J. F. Jardine.
Local persistence: homotopy theory of filtrations.
Oberwolfach Reports, 5(3):1623–1625, 2008.


J.F. Jardine.
Fuzzy sets and presheaves.
Preprint, 2018.


J.F. Jardine.
Stable components and layers.
Preprint, 2019.