Geometric principles of data visualization

Jürgen Jost
Max Planck Institute for Mathematics in the Sciences, Leipzig
ScaDS.AI Dresden/Leipzig, Germany
Santa Fe Institute, New Mexico, USA

We are confronted with ever more and ever larger data sets from many domains, and we want to analyze, interpret and understand them.

We are confronted with ever more and ever larger data sets from many domains, and we want to analyze, interpret and understand them.

These data are often created by elaborate technical devices, like astronomical observations within a wide range frequencies beyond visible light, the traces of collisions of subatomic particles, or by imaging techniques at the molecular level in biology. They are just outside our usual range of sensory experiences and difficult to analyze, interpret and understand.

We are confronted with ever more and ever larger data sets from many domains, and we want to analyze, interpret and understand them.

These data are often created by elaborate technical devices, like astronomical observations within a wide range frequencies beyond visible light, the traces of collisions of subatomic particles, or by imaging techniques at the molecular level in biology. They are just outside our usual range of sensory experiences and difficult to analyze, interpret and understand.

Thus, human psychology may not directly help us

We are confronted with ever more and ever larger data sets from many domains, and we want to analyze, interpret and understand them.

These data are often created by elaborate technical devices, like astronomical observations within a wide range frequencies beyond visible light, the traces of collisions of subatomic particles, or by imaging techniques at the molecular level in biology. They are just outside our usual range of sensory experiences and difficult to analyze, interpret and understand.

Thus, human psychology may not directly help us and we need mathematics!

But mathematics needs structure, and the data are not random.

But mathematics needs structure, and the data are not random. They should therefore carry some structure.

But mathematics needs structure, and the data are not random. They should therefore carry some structure. How to detect or guess such structure?

While the data usually come in some high dimensional spaces, intrinsically, they may be of much lower dimension because of correlations between data points, regularities, invariances, symmetries, ...

But mathematics needs structure, and the data are not random. They should therefore carry some structure. How to detect or guess such structure?

While the data usually come in some high dimensional spaces, intrinsically, they may be of much lower dimension because of correlations between data points, regularities, invariances, symmetries, ...
Mathematical methods should therefore exploit this and find good low dimensional representations that bring out the essential features particular to the data set at hand.

But mathematics needs structure, and the data are not random.They should therefore carry some structure. How to detect or guess such structure?

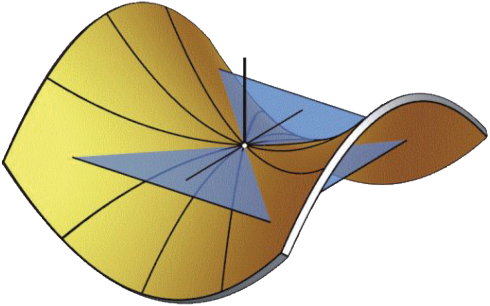
While the data usually come in some high dimensional spaces, intrinsically, they may be of much lower dimension because of correlations between data points, regularities, invariances, symmetries, ...
Mathematical methods should therefore exploit this and find good low dimensional representations that bring out the essential features particular to the data set at hand.
For purposes of visualization, it would even be best if those representations are 2D or 3D.

Methods of Machine Learning

We assume that the data come with some metric structure, that is, distances between data points. These could be Euclidean distances (when the data are presented in some Euclidean space, where the axes record the values of particular quantitative features) or abstract inner distances (dissimilarities). Machine learning then extracts features of the data from these metric relations.

Some methods of Machine Learning

1. Manifold learning
   Data given in some high-dimensional (Euclidean) space, but are assumed to lie on or be concentrated near some low-dimensional smooth submanifold, which may stretch into many ambient directions. This manifold should be recovered by sampling the data.



Graphics courtesy of G. Gläser and K. Polthier

Some methods of Machine Learning

1. Manifold learning
2. Network analysis
   Represent data as graphs, by connecting data points that are sufficiently close or similar or have a particular relation. Use concepts from graph theory to identify particular features.[1]

[1] E.g. M.Eidi, A.Farzam, W.Leal, A.Samal, JJ, *Theory Biosc.*, 2020

Some methods of Machine Learning

1. Manifold learning
2. Network analysis
   Represent data as graphs, by connecting data points that are sufficiently close or similar or have a particular relation. Use concepts from graph theory to identify particular features. More generally, use hypergraphs where more than two elements can entertain a relation, as e.g. in chemical reaction networks.

Some methods of Machine Learning

1. Manifold learning
2. Network analysis
   Represent data as graphs, by connecting data points that are sufficiently close or similar or have a particular relation. Use concepts from graph theory to identify particular features. More generally, use hypergraphs where more than two elements can entertain a relation, as e.g. in chemical reaction networks.
3. Topological data analysis (persistent homology)
   For each radius $r > 0$, turn the intersection pattern of the balls $B(x, r)$ around the data points into a simplicial complex and compute its topological invariants (homology classes). Identify those homology classes that persist for a large range of radii, as encoding important morphological features of the data at their persistence scale.

Machine Learning and Mathematics

1. In manifold learning, a graph is constructed by joining sufficiently close sample points. With more and more samples, eigenvalues/eigenfunctions of graph Laplacian approximate those of Laplace-Beltramie operator of manifold, to recover it.

---

[2] JJ, R.Mulas, D.Zhang, *Spectra of discrete structures*, Cambridge Texts in Advanced Math., to appear

[3] E.g. M.Eidi, JJ, *Sci.Rep.*, 2020; JJ, F.Münch, A.Samal, E.Saucan, *Discrete curvatures and their applications*, monograph, in preparation

[4] P.Joharinad, JJ, *Geometric methods of data analysis*, Mathematics of Data, Springer

Machine Learning and Mathematics

1. In manifold learning, a graph is constructed by joining sufficiently close sample points. With more and more samples, eigenvalues/eigenfunctions of graph Laplacian approximate those of Laplace-Beltramie operator of manifold, to recover it.
2. Network analysis can use eigenvalues and eigenfunctions of graphs and hypergraphs[2] and of curvature statistics,[3] as some foundations of machine learning and data analysis.
3. Intersection patterns of balls in TDA can be related to curvature notions for metric spaces.[4]

[2] JJ, R.Mulas, D.Zhang, *Spectra of discrete structures*, Cambridge Texts in Advanced Math., to appear

[3] E.g. M.Eidi, JJ, *Sci.Rep.*, 2020; JJ, F.Münch, A.Samal, E.Saucan, *Discrete curvatures and their applications*, monograph, in preparation

[4] P.Joharinad, JJ, *Geometric methods of data analysis*, Mathematics of Data, Springer

I shall present a method, **IsUMap**, developed in

L.Barth, H.Fahimi, P.Joharinad, J.J., J.Keck, *Data visualization with category theory and geometry*, Math of Data, Springer, to appear

that builds upon and improves the popular **UMAP**.[5]

[5] McInnes, Healy and Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arxiv, 2018

General aspects

Data are often

- given in some high-dimensional (Euclidean) space

General aspects

Data are often

- given in some high-dimensional (Euclidean) space
- carry distances (metric relations)

General aspects

Data are often

- given in some high-dimensional (Euclidean) space
- carry distances (metric relations)
- possess intrinsic regularities and correlations, that is, are intrinsically lower-dimensional

General aspects

Data are often

- given in some high-dimensional (Euclidean) space
- carry distances (metric relations)
- possess intrinsic regularities and correlations, that is, are intrinsically lower-dimensional
- may therefore possibly sit on a smooth lower-dimensional manifold (this assumption allows for interpolations between data points)

General aspects

Data are often

- given in some high-dimensional (Euclidean) space
- carry distances (metric relations)
- possess intrinsic regularities and correlations, that is, are intrinsically lower-dimensional
- may therefore possibly sit on a smooth lower-dimensional manifold (this assumption allows for interpolations between data points)
- but this manifold need not be linear and may be highly curved and stretch into many ambient directions

General aspects

Data are often

- given in some high-dimensional (Euclidean) space
- carry distances (metric relations)
- possess intrinsic regularities and correlations, that is, are intrinsically lower-dimensional
- may therefore possibly sit on a smooth lower-dimensional manifold (this assumption allows for interpolations between data points)
- but this manifold need not be linear and may be highly curved and stretch into many ambient directions
- and the data need not be uniformly distributed on that manifold

General aspects

Data are often

- given in some high-dimensional (Euclidean) space
- carry distances (metric relations)
- possess intrinsic regularities and correlations, that is, are intrinsically lower-dimensional
- may therefore possibly sit on a smooth lower-dimensional manifold (this assumption allows for interpolations between data points)
- but this manifold need not be linear and may be highly curved and stretch into many ambient directions
- and the data need not be uniformly distributed on that manifold
- and the data should ultimately be visualized in 2 or 3D, preserving the qualitative aspects, like clusters, as well as possible

Existing methods

1. **Isomap:** Tenenbaum, Silve and Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 2000
2. **t-SNE:** Hinton and Roweis, *Stochastic neighbor embedding*, Adv. Neur. Inf. Proc. Sys, 2002
3. **UMAP:** McInnes, Healy and Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arxiv, 2018

General idea:

Connect data points that are sufficiently close to get a graph. When more and more data points are sampled, the graph gets denser and denser and should ultimately approximate a manifold.

General idea:

Connect data points that are sufficiently close to get a graph. When more and more data points are sampled, the graph gets denser and denser and should ultimately approximate a manifold. When we somehow represent, project or approximate those graphs by planar ones, that is, in 2D, we should be able to see the essential features of the data manifold.

General idea:

Connect data points that are sufficiently close to get a graph. When more and more data points are sampled, the graph gets denser and denser and should ultimately approximate a manifold. When we somehow represent, project or approximate those graphs by planar ones, that is, in 2D, we should be able to see the essential features of the data manifold.

Problem: Properties of the data set may only be apparent in high dimension and may get lost by a projection into 2D.

General idea:

Connect data points that are sufficiently close to get a graph. When more and more data points are sampled, the graph gets denser and denser and should ultimately approximate a manifold. When we somehow represent, project or approximate those graphs by planar ones, that is, in 2D, we should be able to see the essential features of the data manifold.

Problem: Properties of the data set may only be apparent in high dimension and may get lost by a projection into 2D.

Approach: Preprocess and modify the graphs to amplify local features, like clusters, before projecting.

The mathematics of UMAP and IsUMap

1. Connect each data point with its k nearest neighbors, to get a collection of star graphs

The mathematics of UMAP and IsUMap

1. Connect each data point with its k nearest neighbors, to get a collection of star graphs
2. Develop a canonical scheme for merging these star graphs, with tools from category theory

The mathematics of UMAP and IsUMap

1. Connect each data point with its k nearest neighbors, to get a collection of star graphs
2. Develop a canonical scheme for merging these star graphs, with tools from category theory
3. Project the resulting space onto 2D with minimal distortion

Riemannian geometry

Bernhard Riemann (1826-1866)

Riemannian geometry

Bernhard Riemann (1826-1866) in 1854 introduced the abstract concept of a manifold of arbitrary dimension n equipped with a metric tensor $(g_{ij})_{i,j=1,\ldots,n}$ with which one can compute a scalar product between tangent vectors at a point.

$$\langle V, W \rangle = \sum_{i,j=1,\ldots,n} g_{ij}(x) v^i w^j$$

for $V = \sum_i v^i \frac{\partial}{\partial x^i}, W = \sum_i w^j \frac{\partial}{\partial x^j}$ at a point x.

[6] J.J., *Bernhard Riemann, On the hypotheses...*, 2nd ed. Birkhäuser, 2025;

Riemannian geometry

Bernhard Riemann (1826-1866) in 1854 introduced the abstract concept of a manifold of arbitrary dimension n equipped with a metric tensor $(g_{ij})_{i,j=1,\dots,n}$ with which one can compute a scalar product between tangent vectors at a point. Such a Riemannian manifold need not be realized in any ambient Euclidean space, but is intrinsically determined by its metric tensor.

$$\langle V, W \rangle = \sum_{i,j=1,\dots,n} g_{ij}(x) v^i w^j \tag{1}$$

for $V = \sum_i v^i \frac{\partial}{\partial x^i}, W = \sum_i w^j \frac{\partial}{\partial x^j}$ at a point x.
But this is represented in local coordinates. In other coordinates, the tangent vectors V, W and the metric tensor (g_{ij}) look different, but they transform in such a way that the scalar quantity (1) remains invariant.
This is the principle of *covariance* that is fundamental for Einstein's *theory of general relativity*. It tells us that we should look for intrinsic properties that do not depend on the representation.

Riemannian geometry

Bernhard Riemann (1826-1866) in 1854 introduced the abstract concept of a manifold of arbitrary dimension n equipped with a metric tensor $(g_{ij})_{i,j=1,\dots,n}$ with which one can compute a scalar product between tangent vectors at a point.
Riemann[6] solved the problem of determining a complete set of local invariants for such a metric tensor. From the *curvature tensor* with components

$$R^k_{\ell ij} = \frac{\partial \Gamma^k_{j\ell}}{\partial x^i} - \frac{\partial \Gamma^k_{i\ell}}{\partial x^j} + \Gamma^k_{im}\Gamma^m_{j\ell} - \Gamma^k_{jm}\Gamma^m_{i\ell} \qquad (1)$$

with $\Gamma^i_{jk} = \frac{1}{2} g^{i\ell}(\frac{\partial}{\partial x^k} g_{j\ell} + \frac{\partial}{\partial x^k} g_{k\ell} - \frac{\partial}{\partial x^\ell} g_{jk})$ (summation signs over double indices omitted; $g^{i\ell}$ is the inverse of the metric tensor), he extracts the **sectional curvatures**.

[6] J.J., *Bernhard Riemann, On the hypotheses...*, 2nd ed. Birkhäuser, 2025; J.J. *Riem.Geom.& Geom.Anal.*, 7th ed. Springer, 2017

Somewhat surprisingly, for the moment, we do not need curvatures, only the concept of a Riemannian metric.

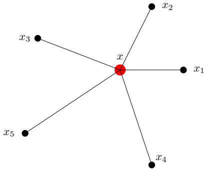
In fact, Riemann introduced another important device, that of *normal coordinates*. Essentially, given a point p in a Riemannian manifold M, for other points q in the vicinity of p, you record the distance $d(p, q)$ and the angles that the shortest geodesic from p to q makes with some reference direction at p. Like euclidean polar coordinates. These coordinates exhibit what the manifold locally looks like from the perspective of p.

In fact, Riemann introduced another important device, that of *normal coordinates*. Essentially, given a point p in a Riemannian manifold M, for other points q in the vicinity of p, you record the distance $d(p, q)$ and the angles that the shortest geodesic from p to q makes with some reference direction at p. Like euclidean polar coordinates.

This does not work globally, because in general the shortest geodesic from p to a more distant point in M need not be unique.

In fact, Riemann introduced another important device, that of *normal coordinates*. Essentially, given a point p in a Riemannian manifold M, for other points q in the vicinity of p, you record the distance $d(p,q)$ and the angles that the shortest geodesic from p to q makes with some reference direction at p. Like euclidean polar coordinates.

Thus, in a discrete metric space, for a point x, we take its k nearest neighbors $x_1, \ldots, x_k$, ranked according to the distance from p and consider the star graph $S(x)$ with x as its central vertex and edges of length $d(x, x_j)$ to the x_j.



In fact, Riemann introduced another important device, that of *normal coordinates*. Essentially, given a point p in a Riemannian manifold M, for other points q in the vicinity of p, you record the distance $d(p,q)$ and the angles that the shortest geodesic from p to q makes with some reference direction at p. Like euclidean polar coordinates.

Thus, in a discrete metric space, for a point x, we take its k nearest neighbors $x_1, \ldots, x_k$, ranked according to the distance from p and consider the star graph $S(x)$ with x as its central vertex and edges of length $d(x, x_j)$ to the x_j.

However, for computational reasons, that metric needs to be modified.

Equip the star graphs with a (generalized) metric

$$\begin{aligned} d_x(x, x_j) &= \frac{d(x, x_j) - \rho_x}{\sigma_x} \quad \text{for } j = 1, \ldots k \\ d_x(y, y) &= 0 \quad \text{for all } y \\ d_x(x_j, x_\ell) &= \frac{1}{\sqrt{2}}(d_x(x_j, x) + d_x(x, x_\ell)) \quad \text{for neighbors } x_j, x_\ell \text{ of } x \\ d_x(y, z) &= \infty \quad \text{in all other cases} \end{aligned} \tag{2}$$

where $d(.,.)$ is the original distance function on the data set.

Equip the star graphs with a (generalized) metric

$$\begin{aligned} d_x(x, x_j) &= \frac{d(x, x_j) - \rho_x}{\sigma_x} \quad \text{for } j = 1, \dots k \\ d_x(y, y) &= 0 \quad \text{for all } y \\ d_x(x_j, x_\ell) &= \frac{1}{\sqrt{2}}(d_x(x_j, x) + d_x(x, x_\ell)) \quad \text{for neighbors } x_j, x_\ell \text{ of } x \\ d_x(y, z) &= \infty \quad \text{in all other cases} \end{aligned} \tag{2}$$

In UMAP, one puts the distance $= \infty$ between different neighbors of the center x, thereby violating the triangle inequality.

Equip the star graphs with a (generalized) metric

$$
\begin{aligned}
d_x(x, x_j) &= \frac{d(x, x_j) - \rho_x}{\sigma_x} \quad \text{for } j = 1, \dots k \qquad (2) \\
d_x(y, y) &= 0 \quad \text{for all } y \\
d_x(x_j, x_\ell) &= \frac{1}{\sqrt{2}}(d_x(x_j, x) + d_x(x, x_\ell)) \quad \text{for neighbors } x_j, x_\ell \text{ of } x \\
d_x(y, z) &= \infty \quad \text{in all other cases}
\end{aligned}
$$

The factor $\frac{1}{\sqrt{2}}$ mimicks the Euclidean distance of the ends of orthogonal vectors. We could also omit it.
The choice of ρ_x and σ_x will be explained in a moment.

Definition (David Spivak)

An *uber-metric space* (X, d) is a set X equipped with a map $d : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

1. $d(x, y) \geq 0$, and $d(x, x) = 0$;
2. $d(x, y) = d(y, x)$; and
3. $d(x, z) \leq d(x, y) + d(y, z)$.

The category of uber-metric spaces $\mathbf{UM}$ has as objects uber-metric spaces and as morphisms again non-expansive maps. The category of finite uber-metric spaces is denoted by $\mathbf{FinUM}$.

The third property in the definition of uber-metric space implies that if $d(x, z) = \infty$ in an uber-metric space, then for any y either $d(x, y) = \infty$ or $d(z, y) = \infty$.

$$
\begin{aligned}
d_x(x, x_j) &= \frac{d(x, x_j) - \rho_x}{\sigma_x} \quad \text{for } j = 1, \ldots k \qquad (3)\\
d_x(y, y) &= 0 \quad \text{for all } y\\
d_x(x_j, x_\ell) &= \frac{1}{\sqrt{2}}(d_x(x_j, x) + d_i(x, x_\ell)) \quad \text{for neighbors } x_j, x_\ell \text{ of } x\\
d_x(y, z) &= \infty \quad \text{in all other cases.}\\
\rho_x &= d(x, x_1)\\
\sigma_x &= d(x, x_k)
\end{aligned}
$$

The choice of ρ_x eases the curse of dimension, because randomly picked points in a high dimensional ball tend to concentrate near the outer boundary. For many data sets, however, already $\rho_x = 0$ works well.

$$
\begin{aligned}
d_x(x, x_j) &= \frac{d(x, x_j) - \rho_x}{\sigma_x} \quad \text{for } j = 1, \dots k \qquad (3)\\
d_x(y, y) &= 0 \quad \text{for all } y\\
d_x(x_j, x_\ell) &= \frac{1}{\sqrt{2}}(d_x(x_j, x) + d_i(x, x_\ell)) \quad \text{for neighbors } x_j, x_\ell \text{ of } x\\
d_x(y, z) &= \infty \quad \text{in all other cases.}\\
\rho_x &= d(x, x_1)\\
\sigma_x &= d(x, x_k)
\end{aligned}
$$

The choice of ρ_x eases the curse of dimension, because randomly picked points in a high dimensional ball tend to concentrate near the outer boundary. For many data sets, however, already $\rho_x = 0$ works well.

σ_x achieves some normalization. This is important if the data points are non-uniformly distributed on the manifold.

$$d_x(x, x_j) = \frac{d(x, x_j) - d(x, x_1)}{d(x, x_k)} \quad \text{for } j = 1, \ldots k \tag{4}$$

$$d_x(x_j, x_\ell) = \frac{1}{\sqrt{2}}(d_x(x_j, x) + d_i(x, x_\ell)) \text{ for neighbors of } x \tag{5}$$

$$d_x(y, z) = \infty \quad \text{in all other cases.} \tag{6}$$

Thus, by (4), the closest neighbor x_1 of x has distance 0, and the largest neighbor x_k has distance < 1. By (5), the vertices of the star graph satisfy a triangle inequality, as if sitting in independent directions. By (6), vertices not in the star have infinite distances and therefore are not seen from the center.

Now, what to do with those local star graphs? How to recombine them into a global structure?

[7] K.Menger, *PNAS*, 1942

Now, what to do with those local star graphs? How to recombine them into a global structure?

Options:

1. Develop a canonical scheme for merging metric spaces
2. Convert them into fuzzy graphs that can be merged in a canonical way with the help of a t-conorm, utilizing constructions from category theory. (Fuzzy graphs are a special case of Spivak's fuzzy simplicial sets. See below)

[7] K.Menger, *PNAS*, 1942

Now, what to do with those local star graphs? How to recombine them into a global structure?

Options:

1. Develop a canonical scheme for merging metric spaces
2. Convert them into fuzzy graphs that can be merged in a canonical way with the help of a t-conorm, utilizing constructions from category theory. (Fuzzy graphs are a special case of Spivak's fuzzy simplicial sets. See below)

I shall develop here the first alternative, but the cognoscenti will see the analogy with Spivak's fuzzy simplicial sets. An essential idea, however, already goes back to Karl Menger.[7]

[7] K.Menger, *PNAS*, 1942

Hazy sets

Definition

Topology on $[0,\infty]$: open sets $=$ intervals $(s,\infty]$, indicated simply by s. $i_{ts} : t \to s$ is inclusion $(t,\infty] \subset (s,\infty]$ for $s \leq t$. Category $\mathbf{H}$.

Hazy sets

Definition

Topology on $[0, \infty]$: open sets = intervals $(s, \infty]$, indicated simply by s. $i_{ts} : t \to s$ is inclusion $(t, \infty] \subset (s, \infty]$ for $s \leq t$. Category $\mathbf{H}$. A *hazy set* S is a sheaf on $\mathbf{H}$ for which all restriction maps $S(i_{ts} : t \to s) : S(s) \to S(t)$ are injections. Their category is denoted by $\mathbf{Haz}$.

Hazy sets

Definition

Topology on $[0, \infty]$: open sets $=$ intervals $(s, \infty]$, indicated simply by s. $i_{ts} : t \to s$ is inclusion $(t, \infty] \subset (s, \infty]$ for $s \leq t$. Category $\mathbf{H}$.
A *hazy set* S is a sheaf on $\mathbf{H}$ for which all restriction maps $S(i_{ts} : t \to s) : S(s) \to S(t)$ are injections. Their category is denoted by $\mathbf{Haz}$.
We call s the *haziness* of $S(s)$ and consider $S(s)$ as the set of haziness at most s.
All $S(s)$ are subsets of $X := S(\infty)$. Set of haziness precisely t

$$S(=t) := S(t) \setminus \bigcup_{s>t} S(s). \tag{7}$$

Hazy sets

Definition

Topology on $[0, \infty]$: open sets $=$ intervals $(s, \infty]$, indicated simply by s. $i_{ts} : t \to s$ is inclusion $(t, \infty] \subset (s, \infty]$ for $s \leq t$. Category $\mathbf{H}$.
A *hazy set* S is a sheaf on $\mathbf{H}$ for which all restriction maps $S(i_{ts} : t \to s) : S(s) \to S(t)$ are injections. Their category is denoted by $\mathbf{Haz}$.
We call s the *haziness* of $S(s)$ and consider $S(s)$ as the set of haziness at most s.
All $S(s)$ are subsets of $X := S(\infty)$. Set of haziness precisely t

$$S(=t) := S(t) \setminus \bigcup_{s>t} S(s). \tag{7}$$

A *hazy simplicial set* is a functor $\mathbf{\Delta^{op}} \to \mathbf{Haz}$ where Δ is the simplicial indexing category. Functor category $\mathbf{sHaz}$ of hazy simplicial sets.

A hazy simplicial set is a functor $S : (\mathbf{\Delta} \times \mathbf{H})^{\mathsf{op}} \to \mathbf{Set}$.
$S^n_s := S([n], (s, \infty])$ are the n-simplices in the image of S of haziness at most s.

Lemma

A simplex is at least as hazy as its faces.

Hazyness becomes metric

Yoneda embedding $(\Delta \times \mathbf{H}) \to \mathbf{Set}^{\Delta \times \mathbf{H}^{op}}$ yields functors Δ_s^n with

$$\Delta_s^n(m, t) = \mathrm{Hom}_{\Delta \times \mathbf{H}}((m, t), (n, s)). \tag{8}$$

Such morphisms exist only for $s \leq t$. And so, morphisms $\Delta_s^n \to \Delta_r^\ell$ (obtained by composing morphisms $((m, t) \to (n, s) \to (\ell, r))$ can exist only for $r \leq s$. We consider Δ_s^n as the n-simplices of haziness at most s.

Hazyness becomes metric

Yoneda embedding $(\Delta \times \mathbf{H}) \to \mathbf{Set}^{\Delta \times \mathbf{H}^{op}}$ yields functors Δ^n_s with

$$\Delta^n_s(m,t) = \mathrm{Hom}_{\boldsymbol{\Delta} \times \mathbf{H}}((m,t),(n,s)). \tag{8}$$

Such morphisms exist only for $s \leq t$. And so, morphisms $\Delta^n_s \to \Delta^\ell_r$ (obtained by composing morphisms $((m,t) \to (n,s) \to (\ell,r))$ can exist only for $r \leq s$. We consider Δ^n_s as the n-simplices of haziness at most s.

The *smear functor* $\mathbf{Sm}$ associates to Δ^n_s a geometric n-simplex of diameter s,

$$\mathbf{Sm}_\Delta(\Delta^n_s) := \left\{ x \in \mathbb{R}^{n+1} \,\middle|\, x^i \geq 0 \text{ and } \sum_{i=1}^{n+1} x^i = s \right\}. \tag{9}$$

On morphisms, it operates by rescaling by $\frac{s}{t}$ when $s \leq t$.

Hazyness becomes metric

Yoneda embedding $(\Delta \times \mathbf{H}) \to \mathbf{Set}^{\Delta\times\mathbf{H}^{op}}$ yields functors Δ^n_s with

$$\Delta^n_s(m,t) = \mathrm{Hom}_{\Delta\times\mathbf{H}}((m,t),(n,s)). \tag{8}$$

Such morphisms exist only for $s \leq t$. And so, morphisms $\Delta^n_s \to \Delta^\ell_r$ (obtained by composing morphisms $((m,t) \to (n,s) \to (\ell,r))$ can exist only for $r \leq s$. We consider Δ^n_s as the n-simplices of haziness at most s.

The *smear functor* $\mathbf{Sm}$ associates to Δ^n_s a geometric n-simplex of diameter s,

$$\mathbf{Sm}_\Delta(\Delta^n_s) := \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; x^i \geq 0 \text{ and } \sum_{i=1}^{n+1} x^i = s \right\}. \tag{9}$$

On morphisms, it operates by rescaling by $\frac{s}{t}$ when $s \leq t$.

Using Kan extensions (while a functor need not admit adjoints, its lift to the Yoneda categories of presheaves does), $\mathbf{Sm}$ becomes a functor from $\mathbf{sHaz}$ to $\mathbf{UM}$.

Hazyness becomes metric

Yoneda embedding $(\Delta \times \mathbf{H}) \to \mathbf{Set}^{\Delta \times \mathbf{H}^{op}}$ yields functors Δ^n_s with

$$\Delta^n_s(m,t) = \mathrm{Hom}_{\Delta \times \mathbf{H}}((m,t),(n,s)). \tag{8}$$

Such morphisms exist only for $s \leq t$. And so, morphisms $\Delta^n_s \to \Delta^\ell_r$ (obtained by composing morphisms $((m,t) \to (n,s) \to (\ell,r))$ can exist only for $r \leq s$. We consider Δ^n_s as the n-simplices of haziness at most s.

The *smear functor* $\mathbf{Sm}$ associates to Δ^n_s a geometric n-simplex of diameter s,

$$\mathbf{Sm}_\Delta(\Delta^n_s) := \left\{ x \in \mathbb{R}^{n+1} \,\middle|\, x^i \geq 0 \text{ and } \sum_{i=1}^{n+1} x^i = s \right\}. \tag{9}$$

On morphisms, it operates by rescaling by $\frac{s}{t}$ when $s \leq t$.

Image of hazy simplicial set is uber-metric: Simplices have metric (9), compatible with glueing along shared faces. Faces cannot be more hazy than simplex itself, hence diameter not larger, shortest path can't be shortened by going into a higher-dimensional simplex.

From an uber-metric space to a simplicial hazy set

We also want to go in the opposite direction.

Definition

The *tightening functor* is

$$\begin{aligned}\mathbf{Ti} : \mathbf{UM} &\to \mathbf{sHaz}\\ Y &\mapsto \mathbf{Ti}(Y) : (\Delta \times \mathbf{H})^{\mathsf{op}} \to \mathbf{Set}\\ &\qquad (n, s) \mapsto \mathrm{Hom}_{\mathbf{UM}}(\mathbf{Sm}(\Delta^n_s), Y)\ .\end{aligned} \tag{10}$$

Morphisms in $\mathbf{UM}$ are distance non-increasing. Therefore, the larger s, the more morphisms there are into a given uber-metric space Y, because $\mathbf{Sm}(\Delta^n_s)$ is a simplex with diameter s.

Theorem

$\mathbf{Ti}$ *is right adjoint to* $\mathbf{Sm}$, *i.e.*

$$\mathrm{Hom}_{\mathbf{UM}}(\mathbf{Sm}(S), Y) \simeq \mathrm{Hom}_{\mathbf{sHaz}}(S, \mathbf{Ti}(Y))\ . \tag{11}$$

The merger tool

Definition

An *m-scheme* is a function $M : [0,\infty] \times [0,\infty] \to [0,\infty]$ with:

(1) Symmetry : $M(s,t) = M(t,s)$,

(2) Monotonicity : $M(s,t) \leq M(v,w)$ if $s \leq v$ and $t \leq w$,

(3) Associativity : $M(r, M(s,t)) = M(M(r,s),t)$,

(4) Boundary condition $M(s,\infty) = s$.

The merger tool

Definition

An *m-scheme* is a function $M : [0,\infty] \times [0,\infty] \to [0,\infty]$ with:

(1) Symmetry : $M(s,t) = M(t,s)$,

(2) Monotonicity : $M(s,t) \leq M(v,w)$ if $s \leq v$ and $t \leq w$,

(3) Associativity : $M(r, M(s,t)) = M(M(r,s),t)$,

(4) Boundary condition $M(s,\infty) = s$.

Any m-scheme satisfies

$$M(s,0) = M(0,s) = 0 \quad \text{for all } s \ . \tag{12}$$

This follows from $M(0,\infty) = 0$, monotonicity and symmetry.

The merger tool

Definition

An *m-scheme* is a function $M : [0,\infty] \times [0,\infty] \to [0,\infty]$ with:

(1) Symmetry : $M(s,t) = M(t,s)$,

(2) Monotonicity : $M(s,t) \leq M(v,w)$ if $s \leq v$ and $t \leq w$,

(3) Associativity : $M(r, M(s,t)) = M(M(r,s),t)$,

(4) Boundary condition $M(s,\infty) = s$.

Extremal examples:

- $M_{\min}(s,t) = \min(s,t)$
- $M_{\text{ext}}(s,t) = \begin{cases} t & \text{if } s = \infty \\ s & \text{if } t = \infty, \\ 0 & \text{else.} \end{cases}$

The merger tool

Definition

An *m-scheme* is a function $M : [0,\infty] \times [0,\infty] \to [0,\infty]$ with:

(1) Symmetry : $M(s,t) = M(t,s)$,

(2) Monotonicity : $M(s,t) \leq M(v,w)$ if $s \leq v$ and $t \leq w$,

(3) Associativity : $M(r, M(s,t)) = M(M(r,s),t)$,

(4) Boundary condition $M(s,\infty) = s$.

$M_{\min}(s,t) = \min(s,t)$, $M_{\text{ext}}(s,t) = 0$ if $s,t \neq \infty$ and (4).
Let $0 \in V \subset \mathbb{R}^{\geq 0}$,

$$M_V(s,t) := \begin{cases} t & \text{if } s = \infty \\ s & \text{if } t = \infty, \\ \sup_{v\in V}\{\ v \mid v \leq s \text{ and } v \leq t\ \} & \text{else.} \end{cases} \tag{12}$$

For $V = \mathbb{R}^{\geq 0}$, we get $M_{\min}$, while for $V = \{0\}$, M_{ext}. If $U \subset V$, then $M_U \leq M_V$. Take $V = [0,a]$ to get a natural family.

Merging

A hazy simplicial set derived from an uber-metric space is a diagram (inverse system) of simplicial sets. From this, we construct a classical hazy simplicial set, which consists of a simplicial set (where $n-$dimensional simplices are identified) along with a haziness function. This hazy simplicial set is then realized as a metric space through a smearing process.

Merging

Consider a finite vertex set V, our data sample, on which a hazy simplicial set is built.

Merging

Properties of a hazy simplicial set enabling metric realization

1- Haziness of simplex $\geq \max$ haziness of faces.

2- Vertices have haziness $= 0$.

3- As scale $\geq$ diameter, entire space is represented as a full simplex equipped with a haziness function.

Merging

Properties of a hazy simplicial set enabling metric realization

1- Haziness of simplex $\geq \max$ haziness of faces.
2- Vertices have haziness $= 0$.
3- As scale $\geq$ diameter, entire space is represented as a full simplex equipped with a haziness function.

In forward direction (uber-metric $\rightarrow$ hazy simplicial set), triangle inequality implicitly reflected in haziness of edges.

Merging

Properties of a hazy simplicial set enabling metric realization

1- Haziness of simplex $\geq \max$ haziness of faces.
2- Vertices have haziness $= 0$.
3- As scale $\geq$ diameter, entire space is represented as a full simplex equipped with a haziness function.

In forward direction (uber-metric $\rightarrow$ hazy simplicial set), triangle inequality implicitly reflected in haziness of edges.

However, in general, edges of a hazy simplicial set need not satisfy triangle inequality. Still, in backward direction (hazy simplicial set $\longrightarrow$ uber-metric), process ensures all uber-metric properties.

Merging

Properties of a hazy simplicial set enabling metric realization

1- Haziness of simplex $\geq \max$ haziness of faces.
2- Vertices have haziness $= 0$.
3- As scale $\geq$ diameter, entire space is represented as a full simplex equipped with a haziness function.

In forward direction (uber-metric $\to$ hazy simplicial set), triangle inequality implicitly reflected in haziness of edges.

However, in general, edges of a hazy simplicial set need not satisfy triangle inequality. Still, in backward direction (hazy simplicial set $\longrightarrow$ uber-metric), process ensures all uber-metric properties.

Example: $\delta_1(x,y) = 1, \delta_1(x,z) = \delta_1(y,z) = 3$ and $\delta_2(x,z) = 1, \delta_2(x,y) = \delta_2(y,z) = 3$. With $M = M_{\min}$, $\delta_M(x,y) = 1, \delta_M(x,z) = 1, \delta_M(y,z) = 3$, violating triangle.

Merging

Properties of a hazy simplicial set enabling metric realization

1- Haziness of simplex $\geq \max$ haziness of faces.
2- Vertices have haziness $= 0$.
3- As scale $\geq$ diameter, entire space is represented as a full simplex equipped with a haziness function.

In forward direction (uber-metric $\to$ hazy simplicial set), triangle inequality implicitly reflected in haziness of edges.

However, in general, edges of a hazy simplicial set need not satisfy triangle inequality. Still, in backward direction (hazy simplicial set $\longrightarrow$ uber-metric), process ensures all uber-metric properties.

Example: $\delta_1(x, y) = 1, \delta_1(x, z) = \delta_1(y, z) = 3$ and $\delta_2(x, z) = 1, \delta_2(x, y) = \delta_2(y, z) = 3$. With $M = M_{\min}$, $\delta_M(x, y) = 1, \delta_M(x, z) = 1, \delta_M(y, z) = 3$, violating triangle.

We therefore need

$$d(x, y) := \inf_{x_o = x, x_1, \ldots, x_\ell = y} \sum_{i=1}^{\ell} \delta(x_{i-1}, x_i)\ . \qquad (13)$$

Star graphs

We had constructed star graphs centered at the sample points x, connecting them with their k nearest neighbors. But when x and y are both among the k nearest neighbors of each other, the weights of the edges in the corresponding star graphs may be different. And other weights are even infinite. So, we use the above scheme to merge these star graphs to reconstruct some graph that represents the entire sample.

Star graphs

We had constructed star graphs centered at the sample points x, connecting them with their k nearest neighbors. But when x and y are both among the k nearest neighbors of each other, the weights of the edges in the corresponding star graphs may be different. And other weights are even infinite. So, we use the above scheme to merge these star graphs to reconstruct some graph that represents the entire sample.

You may ask: Why don't we simply construct a graph from the sample by connecting sufficiently close points and use the original distances? Or connect two points when at least one of them is among the k nearest neighbors of the other?

Star graphs

We had constructed star graphs centered at the sample points x, connecting them with their k nearest neighbors. But when x and y are both among the k nearest neighbors of each other, the weights of the edges in the corresponding star graphs may be different. And other weights are even infinite. So, we use the above scheme to merge these star graphs to reconstruct some graph that represents the entire sample.

You may ask: Why don't we simply construct a graph from the sample by connecting sufficiently close points and use the original distances? Or connect two points when at least one of them is among the k nearest neighbors of the other?
One of the earlier schemes, **Isomap**, essentially does that, but the results become better when we use the apparently more complicated procedure suggested by **UMAP** and refined by us in **IsUMap**. This captures the local structure and the possibly varying density better. The local structure may look different from different points, and this is reflected in the modified star graphs.

Merging star graphs

All star graphs are defined on the same underlying set of sample points, and so we can identify points in different star graphs as $[x]$. When we use the minimum m-scheme, we get

$$d_\sim([x],[x']) = \inf(d_X(p_1,q_1) + \cdots + d_X(p_n,q_n)), \qquad (14)$$

where the infimum is taken over all pairs of sequences $(p_1, \cdots, p_n)$, $(q_1, \cdots, q_n)$ of elements of X, such that

$$p_1 \sim x, \quad q_n \sim x', \quad \text{and} \quad p_{i+1} \sim q_i \ \text{ for all } \ 1 \leq i \leq n-1, \qquad (15)$$

Merging star graphs

All star graphs are defined on the same underlying set of sample points, and so we can identify points in different star graphs as $[x]$. When we use the minimum m-scheme, we get

$$d_{\sim}([x],[x']) = \inf(d_X(p_1,q_1) + \cdots + d_X(p_n,q_n)), \tag{14}$$

where the infimum is taken over all pairs of sequences $(p_1, \cdots, p_n)$, $(q_1, \cdots, q_n)$ of elements of X, such that

$$p_1 \sim x, \quad q_n \sim x', \quad \text{and} \quad p_{i+1} \sim q_i \ \text{ for all } \ 1 \leq i \leq n-1, \tag{15}$$

Other m-schemes yield different results, and we can experiment which one works best for a given data sample.

IsUMap

Given a sample $x_1, \ldots, x_N$ from a metric data set

1. Construct star graphs centered at the x_i with uber-metrics (UM) that assign distance $= \infty$ between different points unless both are among x_i and its the k nearest neighbors. These metrics are normalized to adjust for non-uniform distributions of the samples and the curse of dimensionality.

IsUMap

Given a sample $x_1, \ldots, x_N$ from a metric data set

1. Construct star graphs centered at the x_i with uber-metrics (UM) that assign distance $= \infty$ between different points unless both are among x_i and its the k nearest neighbors. These metrics are normalized to adjust for non-uniform distributions of the samples and the curse of dimensionality.
2. Merge these local star graphs (conceptually by converting them into hazy graphs and use an m-scheme for combining probabilities)

IsUMap

Given a sample $x_1, \ldots, x_N$ from a metric data set

1. Construct star graphs centered at the x_i with uber-metrics (UM) that assign distance $= \infty$ between different points unless both are among x_i and its the k nearest neighbors. These metrics are normalized to adjust for non-uniform distributions of the samples and the curse of dimensionality.
2. Merge these local star graphs (conceptually by converting them into hazy graphs and use an m-scheme for combining probabilities)
3. Use the Dijkstra algorithm to compute the distance function on the resulting global graph (Γ, d_Γ) .

IsUMap

Given a sample $x_1, \ldots, x_N$ from a metric data set

1. Construct star graphs centered at the x_i with uber-metrics (UM) that assign distance $= \infty$ between different points unless both are among x_i and its the k nearest neighbors. These metrics are normalized to adjust for non-uniform distributions of the samples and the curse of dimensionality.
2. Merge these local star graphs (conceptually by converting them into hazy graphs and use an m-scheme for combining probabilities)
3. Use the Dijkstra algorithm to compute the distance function on the resulting global graph (Γ, d_Γ) .
4. Approximate it by a 2D graph (γ, d_γ) using multidimensional scaling, that is, minimizing a function like

$$\sum_{i,j=1,\ldots,N} (d_\gamma(y_i, y_j) - d_\Gamma(x_i, x_j))^2 \tag{16}$$

where $y_i \in \gamma$ corresponds to $x_i \in \Gamma$.

Hazy simplicial sets and TDA

The preceding scheme constructs a hazy simplicial complex from a metric data set.

Hazy simplicial sets and TDA

The preceding scheme constructs a hazy simplicial complex from a metric data set.
A simplicial complex Σ on a vertex set V consists of subsets of V such that whenever $\sigma \in \Sigma$, then also all $\rho \subset \sigma$, the faces of the simplex σ, are in Σ.

Hazy simplicial sets and TDA

The preceding scheme constructs a hazy simplicial complex from a metric data set.
A simplicial complex Σ on a vertex set V consists of subsets of V such that whenever $\sigma \in \Sigma$, then also all $\rho \subset \sigma$, the faces of the simplex σ, are in Σ.
In a hazy simplicial complex, all simplices carry haziness values, and all the faces of σ must have values not larger than σ itself.

Hazy simplicial sets and TDA

The preceding scheme constructs a hazy simplicial complex from a metric data set.
A simplicial complex Σ on a vertex set V consists of subsets of V such that whenever $\sigma \in \Sigma$, then also all $\rho \subset \sigma$, the faces of the simplex σ, are in Σ.
In a hazy simplicial complex, all simplices carry haziness values, and all the faces of σ must have values not larger than σ itself.
In short: The radius r of TDA that determines the intersections of distance balls becomes a haziness value.

Topology and curvature

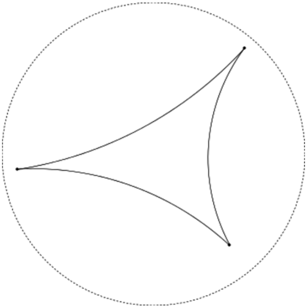
Now, finally, the most important geometric concept, *curvature*, will enter.

Sectional curvature

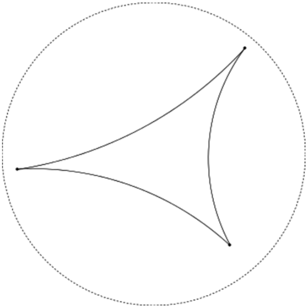
The sectional curvatures, that is, the Riemannian curvatures evaluated on 2-dimensional tangent planes determine the metric locally completely. If they vanish, the manifold is flat, that is, locally Euclidean. The spaces of constant sectional curvatures (spheres for positive, hyperbolic spaces for negative curvature) are the basic model spaces of geometry.

Spaces of constant curvature serve as comparison spaces in geometry. When the curvature of a space satisfies $\lambda \leq K \leq \mu$, its geometry is between those of the spaces of constant curvature λ and μ.

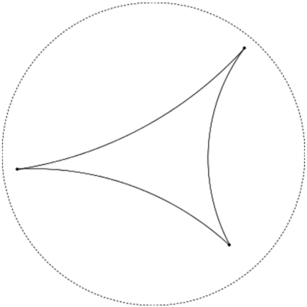
Spaces of constant curvature serve as comparison spaces in geometry. When the curvature of a space satisfies $\lambda \leq K \leq \mu$, its geometry is between those of the spaces of constant curvature λ and μ.
We can, for instance, compare the geometry of spaces of curvature ≤ 0 with that of Euclidean spaces.



A triangle in a space of negative curvature is thinner than in Euclidean space.



A triangle in a space of negative curvature is thinner than in Euclidean space. And it becomes the thinner, the more negative the curvature, and ultimately, when curvature $\to -\infty$, converges to a tripod.



A triangle in a space of negative curvature is thinner than in Euclidean space.
In fact, this is a local property that is characteristic for negative curvature. But such a property can also be checked in metric spaces more general than Riemannian manifolds. We only need geodesic triangles, that is, triangles whose sides are shortest geodesics.

It was then an important idea of Karl Menger (1902-1985), Abraham Wald (1902-1950), Alexandr Danilovic Alexandrov (1912-1999) and Herbert Busemann (1905-1994) to define curvature bounds for metric spaces more general than Riemannian manifolds in terms of inequalities for geodesic triangles.

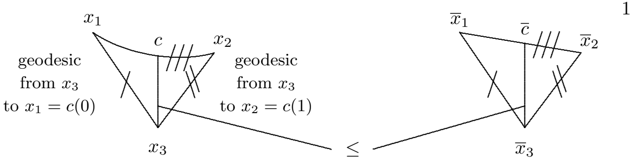


Figure 1: Comparison between a triangle in a space of nonpositive curvature in the sense of Alexandrov and the triangle with the same lengths of corresponding sides (indicated by slashes) in the Euclidean plane.

It was then an important idea of Karl Menger (1902-1985), Abraham Wald (1902-1950), Alexandr Danilovic Alexandrov (1912-1999) and Herbert Busemann (1905-1994) to define curvature bounds for metric spaces more general than Riemannian manifolds in terms of inequalities for geodesic triangles.

But our data spaces are discrete, and therefore do not contain geodesic connections between data points. Therefore, we need a still more general concept.

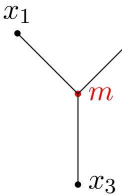
Sectional curvature

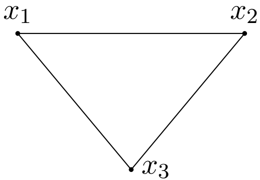
With Parvaneh Joharinad, I have developed a notion of sectional curvature that applies to general metric spaces and can be applied in data analysis.

Here is the basic question:

Sectional curvature

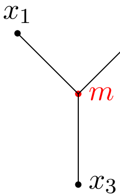
How to distinguish a tripod from a triangle?

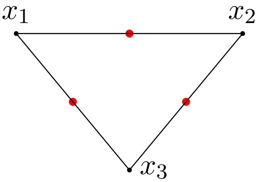




Sectional curvature

How to distinguish a tripod from a triangle?

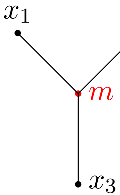


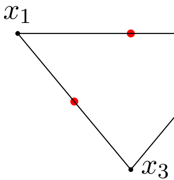


Let $d(x_i, x_j) = 2r$ for $i \neq j$. Then the balls $B(x_i, r) = \{x : d(x_i, x) \leq r\}$ intersect pairwise in each graph.

Sectional curvature

How to distinguish a tripod from a triangle?





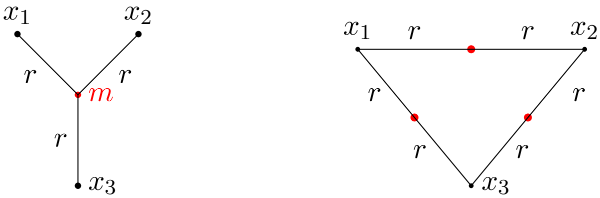
Let $d(x_i, x_j) = 2r$ for $i \neq j$. Then the balls $B(x_i, r) = \{x : d(x_i, x) \leq r\}$ intersect pairwise in each graph. But in the tripod, they also have a triple intersection

$$B(x_1, r) \cap B(x_2, r) \cap B(x_3, r) = \{m\} \neq \emptyset$$

whereas in the triangle, only

$$B(x_1, 2r) \cap B(x_2, 2r) \cap B(x_3, 2r) \neq \emptyset \ .$$

Sectional curvature



$$B(x_i, r) \cap B(x_j, r) \neq \emptyset$$

In the tripod, they also have a triple intersection

$$B(x_1, r) \cap B(x_2, r) \cap B(x_3, r) = \{m\} \neq \emptyset$$

whereas in the triangle, only

$$B(x_1, 2r) \cap B(x_2, 2r) \cap B(x_3, 2r) \neq \emptyset \ .$$

These are extreme cases, and when $B(x_i, r) \cap B(x_j, r) \neq \emptyset$, there exists some $1 \leq \lambda \leq 2$ with

$$B(x_1, \lambda r) \cap B(x_2, \lambda r) \cap B(x_3, \lambda r) \neq \emptyset$$

Sectional curvature

$B(x_i, r) \cap B(x_j, r) \neq \emptyset$, $1 \leq \lambda \leq 2$ with

$$B(x_1, \lambda r) \cap B(x_2, \lambda r) \cap B(x_3, \lambda r) \neq \emptyset$$

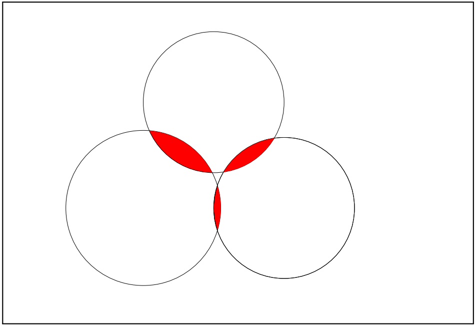
quantifies the curvature. The smaller λ, the more negative the curvature is.

Sectional curvature

$B(x_i, r) \cap B(x_j, r) \neq \emptyset$, $1 \leq \lambda \leq 2$ with

$$B(x_1, \lambda r) \cap B(x_2, \lambda r) \cap B(x_3, \lambda r) \neq \emptyset$$

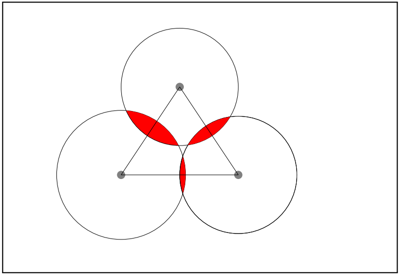
quantifies the curvature. The smaller λ, the more negative the curvature is.



Topological Data Analysis (TDA): For metric family $(x_i)_{i \in I}, d$ and $r > 0$, the Čech complex contains a q-simplex whenever

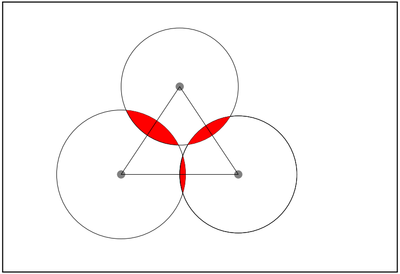
$$\bigcap_{i=1,\dots,q+1} B(x_i, r) \neq \emptyset.$$

Its homology (unfilled simplices) varies as a function of r. Thus, the vertices (0-simplices) of our simplicial complex correspond to the balls. Two vertices are connected by an edge (1-simplex) when the balls intersect, and we fill a triangle (2-simplex) when the three balls have a common intersection.



Here, triangle is not filled, because no triple intersection.

Topological Data Analysis (TDA): The vertices (0-simplices) of our simplicial complex correspond to the balls. Two vertices are connected by an edge (1-simplex) when the balls intersect, and we fill a triangle (2-simplex) when the three balls have a common intersection.

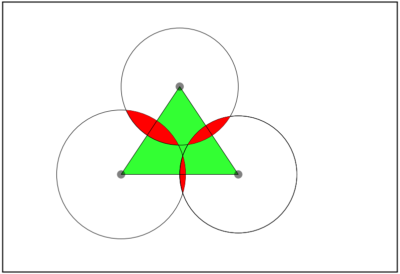


Here, triangle is not filled, because no triple intersection. If the balls were smaller, they would not intersect pairwise, and if they were larger, we would also have a triple intersection. Thus, here we have a scale with non-trivial homology.

Topological Data Analysis (TDA): For metric family $(x_i)_{i\in I}, d$ and $r > 0$, the Čech complex contains a q-simplex whenever

$$\bigcap_{i=1,\ldots,q+1} B(x_i, r) \neq \emptyset.$$

Its homology (unfilled simplices) varies as a function of r.
In contrast, in the Vietoris-Rips complex, simplices are filled whenever the balls around their vertices intersect pairwise.



Our notion of sectional curvature quantifies the difference between the Čech complex and the Vietoris-Rips complex.

How much does one have to enlarge balls that intersect pairwise to get triple intersections?

Our notion of sectional curvature quantifies the difference between the Čech complex and the Vietoris-Rips complex.

How much does one have to enlarge balls that intersect pairwise to get triple intersections?
The less one has to enlarge them, the more negative the curvature.
Thus, we can understand TDA from a geometric perspective.

The indicated notion of sectional curvature for metric spaces[8] is useful for the large scale analysis of metric spaces. When the space is discrete, one may allow for some $\delta > 0$, depending on the scale of the metric and ask for intersections of radius $r + \delta$.

[8] P.Joharinad, J.J. *Mathematical principles of topological and geometric data analysis*, Math of Data, Springer, 2023

The indicated notion of sectional curvature for metric spaces[8] is useful for the large scale analysis of metric spaces. When the space is discrete, one may allow for some $\delta > 0$, depending on the scale of the metric and ask for intersections of radius $r + \delta$.
For instance, for a traffic network, it is useful to distinguish a large scale tripod type pattern, with a center to which all lines go, like the French railway system with Paris as the center, from a ring type structure where locations typically lie on a cycle.

[8] P.Joharinad, J.J. *Mathematical principles of topological and geometric data analysis*, Math of Data, Springer, 2023

Curvature and convexity

The more negative the curvature, the more convex the space becomes.[9]

A region R is *convex* if whenever x and y are in R, then also any point between x and y is in R as well.

[9] J.J., *Nonpositive curvature*, Birkhäuser, 1997

The geometry of conceptual spaces

- *Adjectives* are represented by *convex* regions in conceptual spaces (P.Gärdenfors, *The Geometry of Meaning*, MIT Press, 2014).
  Clustering algorithms also typically, but not necessarily, yield convex regions.

The geometry of conceptual spaces

- *Adjectives* are represented by *convex* regions in conceptual spaces (P.Gärdenfors, *The Geometry of Meaning*, MIT Press, 2014).
  In particular, the regions for the various colors (green, red, blue, yellow,...) in color space are convex.

The geometry of conceptual spaces

- *Adjectives* are represented by *convex* regions in conceptual spaces (P.Gärdenfors, *The Geometry of Meaning*, MIT Press, 2014).
- *Verb* meanings have also the properties of *monotonicity* (larger efforts lead to larger results) and *continuity/discontinuity* (small increases of effort can lead to small/large results) (P.Gärdenfors, M.Warglien, J.J., in: *Frontiers in Psychology*, 2018)

The geometry of conceptual spaces

- *Adjectives* are represented by *convex* regions in conceptual spaces (P.Gärdenfors, *The Geometry of Meaning*, MIT Press, 2014).
- *Verb* meanings have also the properties of *monotonicity* (larger efforts lead to larger results) and *continuity/discontinuity* (small increases of effort can lead to small/large results) (P.Gärdenfors, M.Warglien, J.J., in: *Frontiers in Psychology*, 2018)
- *Nouns* have a more abstract geometric representation in terms of local sections of presheaves (P.Gärdenfors, M.Warglien, J.J., under review)

Literature

P.Joharinad, J.Jost Math of Data, Springer, 2023

Mathematical principles of topological and geometric data analysis, Math of Data, Springer, 2023

L.Barth, F.Fahimi, P.Joharinad, J.Jost, J.Keck,

Data visualization with category theory and geometry, Math of Data, Springer, to appear

L.Barth, F.Fahimi, P.Joharinad, J.Jost, J.Keck,

IsUMap: Manifold Learning and Data Visualization..., arXiv:2407.17835, with code

L.Barth, F.Fahimi, P.Joharinad, J.Jost, J.Keck,

Fuzzy simplicial sets and their application ..., arXiv:2406.11154

L.Barth, F.Fahimi, P.Joharinad, J.Jost, J.Keck,

Merging Hazy Sets with m-Schemes: A Geometric Approach to Data Visualization, arXiv:2503.01664