Expansion of the theory of metric spaces and fuzzy simplicial sets and their applications to data analysis

Lukas Barth

Joint work with: Fatemeh Fahimi, Parvaneh Joharinad, Jürgen Jost, Janis Keck and Thomas Jan Mikhail

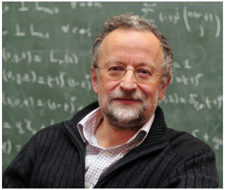
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

ACT, June 2024

Collaborators

Thomas Jan Mikhail

Janis Keck



Jürgen Jost

Parvaneh Joharinad

Fatemeh (Hannaneh) Fahimi

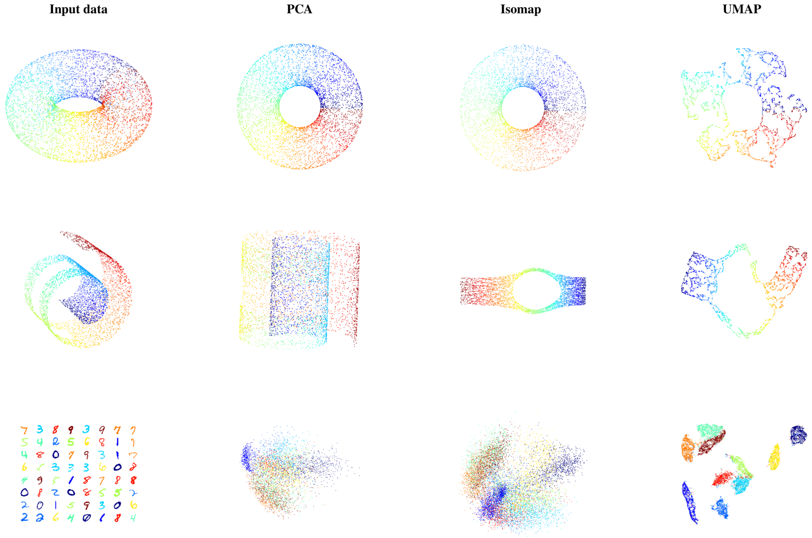
Contents

Contents

Introduction

- How to apply category theory to manifold learning and data visualization?
- Suppose $X \subset \mathbb{R}^n$ is a finite dataset. "Manifold learning" is about extracting a manifold of dimension $d \ll n$, around which the dataset is concentrated.
- Usually by generating $\mathbb{R}^d$ embeddings that can be thought of as local charts or coordinates of the manifold.
- Those embeddings can help to interpolate the data, increase the computational efficiency of downstream tasks and, when $d = 2$ or $d = 3$, serve as visualization.

Introduction



Contents

Metric spaces

- An *uber-metric space* (X, d) is a set X with a map $d : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ s.t.
  1. $d(x, y) \geq 0$, and $d(x, y) = 0$ only if $x = y$;
  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 non-expansive maps, i.e. $d_Y(f(x), f(x')) \leq d_X(x, x')$.
- One can split a metric space with N points into N metric spaces $\{(X, d_i)\}_{i \in \{1, \cdots, N\}}$, where d_i is a "nearest-neighborhood metric":

$$
\begin{aligned}
d_i(x_i, x_{i_j}) = f_i(d(x_i, x_{i_j})) \quad \text{for } j = 1, \dots k \\
d_i(x, x) = 0 \quad \text{for all } x \in X, \qquad \text{and} \qquad d_i(x_j, x) = \infty \quad \text{else.}
\end{aligned}
\tag{1}
$$

  where x_{i_j} is the j-th neighbor of $x_i \in X$.
- The idea is that the finite distances in those neighbourhoods are close to the ones on the underlying manifold around which one assumes the data distribution to be concentrated
- But how to combine those neighborhoods?

Fuzzy simplicial sets

- A *fuzzy set* S is a sheaf on $\mathbf{I} = [0,1]$ for which all restriction maps $S(i_{ab} : a \to b) : S(b) \to S(a)$ are injections. Their category is denoted by $\mathbf{Fuz}$.
- A *classical fuzzy set* is a pair (X, η) where X is a set and $\eta : X \to [0,1]$ is a function, called *strength function*.
  Morphisms in $\mathbf{cFuz}$ are functions $f : (X, \eta) \to (Y, \xi)$ such that $\xi(f(x)) \geq \eta(x) \; \forall x \in X$.
- Fuzzy sets and classical fuzzy sets are isomorphic.
- Δ denotes the *simplicial indexing category*. Its objects are given by finite totally ordered sets $[n] := \{0, 1, \ldots, n\}$ with exactly $n+1$ elements and its morphisms are order preserving maps ($f : [n] \to [m]$ s.t. $f(a) \geq f(b)$ if $a \geq b$).
- A *fuzzy simplicial set* is simply a functor $\Delta^{\mathrm{op}} \to \mathbf{Fuz}$.
  One can also think of them as functors $(\Delta \times \mathbf{I})^{\mathrm{op}} \to \mathbf{Sets}$.
  Their category (morphisms are natural transformations) is denoted by $\mathbf{sFuz}$.[1]
- Think of a simplicial set, where every simplex has a strength.
  And the strength of a simplex is $\leq$ than the minimum of the strength of its faces.

[1] They were introduced by David Spivak.

Contents

- David Spivak showed that there exists an adjunction between $\mathbf{UM}$ and $\mathbf{sFuz}$.

$$\mathrm{Re}_\Delta : \mathbf{y}(\Delta \times \mathbf{I}) \to \mathbf{UM}, \quad \mathrm{Re}_\Delta(\mathbf{y}(n,a)) := \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; \sum_{i=1}^{n+1} x_i = -\log(a) \right\},$$

- As $\mathbf{UM}$ has small colimits, this can be Kan-extended to

$$\begin{aligned} \mathrm{Re} &: \mathbf{sFuz} \to \mathbf{UM}, \\ \mathrm{Re}(S) &:= \mathrm{colim}(D_S) \\ \text{where} \quad D_S &= \mathrm{Re}_\Delta \circ \mathbf{y} \circ P_S : \mathbf{El}(S) \to \mathbf{UM}. \end{aligned} \tag{2}$$

- UMAP uses a similar adjunction $\mathrm{Re}^{\mathrm{U}} : \mathbf{Fin\text{-}sFuz} \to \mathbf{FinEPMet}$:

$$\begin{aligned} \mathrm{Re}^{\mathrm{U}}_\Delta(\mathbf{y}(n,a)) &:= (\{x_0, \cdots, x_n\}, d_a), \\ \text{where} \quad d_a(x_i, x_j) &:= \begin{cases} -\log(a), & \text{if } i \neq j, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{3}$$

Adjunctions

- Suppose that $\text{Re}_\triangle : \mathbf{y}(\Delta \times \mathbf{I}) \to \mathbf{C}$ is any functor and that $\mathbf{C}$ has small colimits. Then the following defines a functor:

$$\begin{aligned} &\text{Re} : \ \mathbf{sFuz} \to \mathbf{C}, \qquad \text{Re}(S) := \text{colim}(D_S) \\ &\text{where} \quad D_S = \text{Re}_\triangle \circ \mathbf{y} \circ P_S : \mathbf{El}(S) \to \mathbf{C}. \end{aligned} \tag{4}$$

  and its right adjoint is

$$\text{Sing}(Y)(n,a) := \text{Hom}_{\mathbf{C}}(\text{Re}_\triangle(\mathbf{y}(n,a)), Y). \tag{5}$$

- $\mathbf{UM}$ and $\mathbf{EPMet}$ both have small colimits. Hence there are infinitely many adjunctions of the types discovered by D. Spivak and the authors of UMAP.
- One can show that $\text{Sing}^{\text{U}}(X,d)(n,a)$ is equivalent to tuples $[x_0, \cdots, x_n] \in X^{n+1}$ with strength at least a, which turns out to be a rescaled **Vietoris-Rips** complex!
- However, the colimit in (4) might be hard to compute.

- UMAP corresponds to the upper right path of the following diagram:

$$
\begin{tikzcd}
\mathbf{Met} \arrow[r, "\text{split}"] & \mathbf{UM}_*^N \arrow[r, "\text{Sing}^N"] \arrow[d, dashed, "\text{merge}_{\mathbf{UM}}"'] & \mathbf{sFuz}_*^N \arrow[r, "\text{ctr}_1^N"] & \mathbf{c1Fuz}_*^N \arrow[r, "\text{merge}_{\mathbf{c1Fuz}}"] & \mathbf{c1Fuz} \arrow[d, "\text{embedding}"'] \\
& \mathbf{UM} \arrow[rrr, dashed, "\text{embedding}"] & & & \mathbf{Euc}
\end{tikzcd}
$$

- The embedding is obtained by minimizing the objective (fuzzy cross-entropy)

$$\mathcal{L}(\{\mathbf{x}\}) := -\sum_{i,j}\{G_{ij}\log(H(\{\mathbf{x}\})_{ij}) + (1-G_{ij})\log(1-H(\{\mathbf{x}\})_{ij})\} \tag{6}$$

  where G is the graph obtained from the high-dim dataset with N points and $H(\{\mathbf{x}\})$ is a graph, obtained from the distances between N vectors $\mathbf{x} \in \mathbb{R}^d$.

- What if one makes use of the full adjunction, computing an explicit description of $\text{merge}_{\mathbf{UM}} := \text{Re} \circ \text{merge}_{\mathbf{sFuz}} \circ \text{Sing}^N$ and uses a geometric embedding instead?

Contents

The colimit in **UM** and **EPMet**

- The colimit of a small diagram $D : \mathbf{I} \to \mathbf{UM}$ is given by

$$\text{colim}(D) = (\text{colim}(FD),\ d_{\text{colim}}) \tag{7}$$

where $\text{colim}(FD)$ is the usual colimit in $\mathbf{Sets}$, while d_{colim} is defined by

$$d_{\sim}([x],[x']) = \inf(d_X(p_1,q_1) + \cdots + d_X(p_n,q_n)), \tag{8}$$

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

$$p_1 \sim x, \quad q_n \sim x', \quad \text{and} \quad p_{i+1} \sim q_i \ \text{ for all } \ 1 \leq i \leq n-1, \tag{9}$$

and d_X is defined by

$$d_X(p_i,q_i) := \begin{cases} d_J(p_i,q_i), & \text{if } p_i, q_i \in FD(J) \\ \infty, & \text{else.} \end{cases} \tag{10}$$

- We have $\mathsf{Re}^{\mathsf{U}}(S) \simeq \mathsf{Re}_{c1}(C_1(\mathsf{tr}_1(S)))$, where:
- $\mathsf{Re}_{c1} : \mathbf{c1Fuz} \to \mathbf{UM}$ is defined by $\mathsf{Re}_{c1}(S) := (S_0, d)$, where $S_0 \xleftarrow{S(\delta_1)} S_1 \xrightarrow{S(\delta_2)} S_0$ is a classical fuzzy graph, and

$$d(x,y) := \inf_{x=x_1,\cdots,x_n=y} \sum_{i=1}^{n-1} d_{\min}(x_i, x_{i+1}), \tag{11}$$

where $d_{\mathsf{min}}(x_1, x_2) := \min\{-\log(\xi_1(s)) \mid [x_1, x_2] \simeq s \in S_1\}$.

- Intuitively: $\mathsf{Re}^{\mathsf{U}}(S)$ generates a metric space, in which the distances are geodesic "graph-hopping" distances, along edges of $\mathsf{tr}_1(S)$
- We used this to show: $\mathsf{Re}^{\mathsf{U}} \circ \mathsf{Sing}^{\mathsf{U}} \simeq \mathsf{id}_{\mathbf{C}}$, where $\mathbf{C}$ is either $\mathbf{UM}$ or $\mathbf{EPMet}$.

Contents

- A *t-conorm* is a function $T : [0, 1] \times [0, 1] \to [0, 1]$ that fulfills some axioms.
- Given a t-conorm T and two classical fuzzy sets (A, ξ_1) and (A, ξ_2), with the same underlying set A, define $\text{merge}_{\mathbf{cFuz}} : \mathbf{cFuz} \times_{\mathbf{Sets}} \mathbf{cFuz} \to \mathbf{cFuz}$ by

$$\begin{aligned} \text{merge}_{\mathbf{cFuz}}((A, \xi_1), (A, \xi_2)) &:= (A, \xi), \\ \text{where} \quad \xi(a) &:= T(\xi_1(a), \xi_2(a)). \end{aligned} \tag{12}$$

- The isomorphism $C : \mathbf{Fuz} \to \mathbf{cFuz}$ then gives us $\text{merge}_{\mathbf{Fuz}}$,
- which in turn yields $\text{merge}_{\mathbf{sFuz}} : \mathbf{sFuz} \times_{\mathbf{sSet}} \mathbf{sFuz} \to \mathbf{sFuz}$ via

$$\text{merge}_{\mathbf{sFuz}}(S_1, S_2)(n, a) := \text{merge}_{\mathbf{Fuz}}(S_1(n, -), S_2(n, -))(a). \tag{13}$$

- We proved that this is indeed a well-defined functor.

The merge in **UM**

- We can finally describe $\text{merge}_{\mathbf{UM}} := \text{Re}^{\text{U}} \circ \text{merge}_{\mathbf{sFuz}} \circ \text{Sing}^{\text{U}}$:
- The functor

$$\text{merge}_{\mathbf{UM}} := \text{Re} \circ \text{merge}_{\mathbf{sFuz}} \circ \text{Sing}^{N} : \mathbf{UM} \times_{\mathbf{Sets}} \cdots \times_{\mathbf{Sets}} \mathbf{UM} \to \mathbf{UM}. \tag{14}$$

can be given the following explicit description:

$$\begin{aligned} &\text{merge}_{\mathbf{UM}}((X, d_1), \cdots, (X, d_N)) = (X, d), \text{ where} \\ &d(x, y) := \inf_{x = x_1, \cdots, x_n = y} \sum_{i=1}^{n-1} (-\log(T_R(x_i, x_{i+1}))), \end{aligned} \tag{15}$$

where T_R is defined recursively in terms of a chosen t-conorm T.

- Combining this with a metric embedding method like classical or (non-)metric multidimensional scaling (MDS) yields a new dimension reduction algorithm.

Contents

Concise description of the method

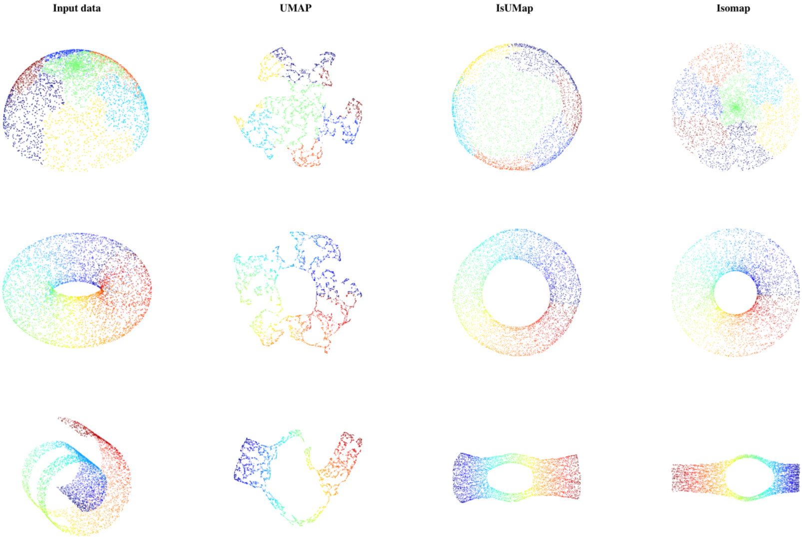
Input: $X \subset \mathbb{R}^n,\ |X| < \infty,\ k \in \mathbb{N},\ m \leq n$.

1. Split $X \subset \mathbb{R}^n$ into $N := |X|$ metric spaces (X, d_i), where d_i is defined by (1).
2. Apply the merge functor defined on the last slide, to obtain the metric space (X, d). One can use Dijkstra's algorithm to compute the infimum.
3. Embed (X, d) into $\mathbb{R}^m$ using classical or (non-)metric multidimensional scaling.

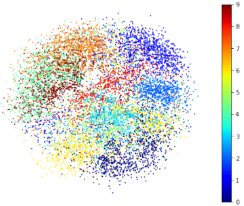
Output: $Y \subset \mathbb{R}^m,\ |Y| = |X|$.

- Similar to UMAP because we proved that $\mathrm{Re}^{\mathsf{U}} \circ \mathrm{Sing}^{\mathsf{U}} \simeq \mathrm{id}_{\mathbf{UM}}$ and we used $\mathrm{merge}_{\mathbf{UM}} := \mathrm{Re}^{\mathsf{U}} \circ \mathrm{merge}_{\mathbf{sFuz}} \circ \mathrm{Sing}^{\mathsf{U}}$.
- At the same time, can yield Isomap as special case, while adding the capabilities to use arbitrary t-conorms, non-classical metric MDS and a uniformization of the data distribution.
- Since it combines UMAP and Isomap and takes place entirely in the category $\mathbf{UM}$, we call our method **IsUMap**.

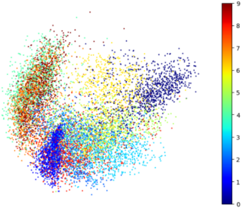
Simulation results



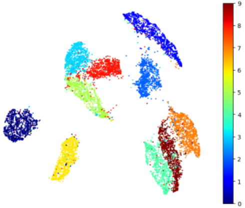
Simulation results



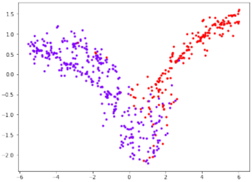
(a)



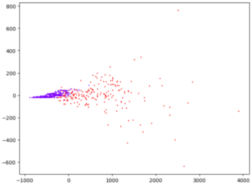
(b)



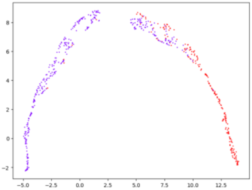
(c)



(d)



(e)



(f)

Contents

Outlook

- One could combine our method with Functorial manifold learning and Functorial clustering via simplicial complexes as introduced by Dan Shiebler.
- As remarked in On UMAP's true loss function by Damrich and Hamprecht, the distortion in UMAP's embedding is largely an effect of negative undersampling, that is not captured by the formal theory describing UMAP. Hence, more effort is needed to understand this effect in categorical terms, possibly by looking at it through the "lens" of Backprop as a functor by Fong et al., or extensions of Learners language by David Spivak, or ideas from Categorical systems theory by David Jaz Myers or using Categorical cybernetics by Capucci, Gavranovic, Hedges and Rischel, and others.
- There is also an interesting connection to TDA because $\mathrm{Sing}^{\mathrm{U}}$ is closely related to the Vietoris-Rips filtration, while objects in $\mathbf{sFuz}$ can also capture geometric (as opposed to only topological) information.

References and contact

- Our preprint: https://arxiv.org/abs/2406.11154
  “Fuzzy simplicial sets and their application to geometric data analysis”
- Our code: https://github.com/LUK4S-B/IsUMap
- Contact me anytime: lukas.barth@mis.mpg.de