Connecting hidden dots: applying topological data analysis to a socioeconomic dataset

Siyuan Zhang

Supervised by Paul Fijn and TriThang Tran

THE UNIVERSITY OF MELBOURNE

Introduction

Topological data analysis (TDA) is a field of study that provides tools for describing the shape of data at multiple scales. In particular, we consider TDA as the process of assigning to a dataset, a family of **simplicial complexes**, and computing **persistent homology**. We apply this to a socioeconomic dataset, and produce interpretable results.

Simplicial complexes

The **k-simplex** spanned by a set of **affinely independent** points $\{x_0, \ldots, x_k\}$ is the set of all points

$$z = \sum_{i=0}^{k} a_i x_i, \qquad \sum_{i=0}^{k} a_i = 1$$

The **interior** of a simplex S, denoted $\mathrm{int}(S)$, is the subset of points where $a_i > 0$ for all i. The **boundary** of S is $S \setminus \mathrm{int}(S)$. A **face** of S is any simplex spanned by a subset of its points.

Figure 1. Standard simplices with interiors drawn in blue [3]

Simplices can be glued together along their faces to form **simplicial complexes**, which are discrete geometric objects that enable us to model data as topological spaces. They are simple enough for algorithmic purposes but retain enough information about the data.

Example

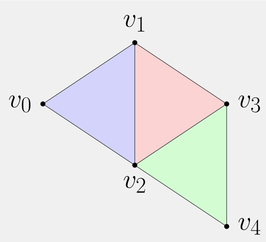


Figure 2. A simplicial complex X consisting of three 2-simplices

This is a valid simplicial complex, since every face (a vertex or an edge)of a simplex in X is itself a simplex in X. The simplices are glued along the edges $\{v_1, v_2\}$ and $\{v_2, v_3\}$, which are faces of each simplex they touch.

Simplicial homology

To analyse the topological features of simplicial complexes in a computationally tractable way, we need the notion of simplicial homology.

For a simplicial complex X, the **k-chains** $C_k(X; \mathbb{F})$ is the $\mathbb{F}$-vector space with the set of **oriented** k-simplices as the basis. Note that an orientation on a simplicial complex can be induced as an ordering on its vertices.

Using the idea of k-chains, we then define the **boundary map** $\partial_k : C_k(X; \mathbb{F}) \to C_{k-1}(X; \mathbb{F})$, which algebraically encodes the boundary of a simplex. More precisely, it is a linear transformation specified as

$$\partial_n([v_0, \ldots, v_k]) \mapsto \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v_i}, \ldots, v_k]$$

where the $\hat{v_i}$ notation means we delete the vertex.

The ***k*th homology group** $H_k(X; \mathbb{F})$ is the quotient group $\ker(\partial_k)/\mathrm{im}(\partial_{k+1})$. We can think of homology groups as the set of cycles in $C_k(X; \mathbb{F})$ that are not the boundaries of elements of $C_{k+1}(X; \mathbb{F})$, i.e. the k+1-simplex the cycle is supposed to enclose is missing.

- H_0 measures the number of path components in a simplicial complex
- H_k measures the number of *k*-dimensional geometric features ("holes") in the complex

In this poster, we work with the field $\mathbb{R}$.

Example

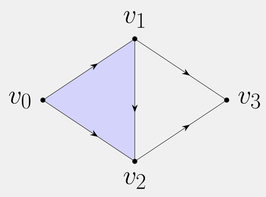


Figure 3. A simplicial complex X consisting of one 2-simplex and one hollow 2-simplex

X has one connected path component and one 1-dimensional hole. In particular, for the 1-dimensional cycles we have

$$\begin{cases} [v_0, v_1] + [v_1, v_2] - [v_0, v_2] \in \ker(\partial_1) \\ [v_1, v_2] + [v_2, v_3] - [v_1, v_3] \in \ker(\partial_1) \\ [v_0, v_1] + [v_1, v_2] - [v_0, v_2] \in \mathrm{im}(\partial_2) \\ [v_1, v_2] + [v_2, v_3] - [v_1, v_3] \notin \mathrm{im}(\partial_2) \end{cases}$$

Therefore, we have $\ker(\partial_1) \cong \mathbb{R} \oplus \mathbb{R}$, $\mathrm{im}(\partial_2) \cong \mathbb{R}$ and the following:

$$\begin{cases} H_0(X; \mathbb{R}) \cong \mathbb{R}, \\ H_1(X; \mathbb{R}) = \ker(\partial_1)/\mathrm{im}(\partial_2) \cong \mathbb{R}, \\ H_k(X; \mathbb{R}) = 0, k > 1 \end{cases}$$

Persistent homology

We can construct a simplicial complex on a finite metric space (X, ∂_X) built from a dataset. For such a metric space, the **Vietoris-Rips complex** $\mathrm{VR}_\epsilon(X, \partial_X)$ is the abstract simplicial complex with

- vertices being the points of X
- a k-simplex $[v_0, v_1, \ldots, v_k]$ when $\partial_X(v_i, v_j) \leq \epsilon \quad \forall 0 \leq i, j \leq k$. (some definitions use 2ϵ)

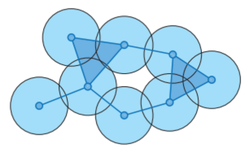


Figure 4. A Vietoris-Rips complex with 2ϵ criterion. [3]

The topological **features**, $\gamma \in H_k\left(\mathrm{VR}_{\epsilon_i}(X; \partial_X)\right)$ change as ϵ varies.

Since the **feature scale** of the data is typically unknown a priori, we choose a range of values of ϵ and look at features that persist over large ranges.

For a feature γ, we say it is:

- **born** at ϵ_i if it does not exist for all $\epsilon' < \epsilon_i$, and
- **dies** at ϵ_j if it becomes zero at ϵ_j or if its image coincides with the image of another feature that was born earlier.

We are interested in persistent features, or features that have a large birth-death interval, since they are more likely to reflect meaningful topological structures in the data.

Methodology

Results in this poster come from a subset of a dataset on the impact of **unconditional cash transfers** (UCTs) to poor households in Kenya. Cash was given either as a lump sum or in monthly transfers, and the effects it had on various socioeconomic indices were measured. [2]

The dataset is cleaned of rows with missing values and ran through the program Ripser to obtain a **persistence diagram**. [1]

A **persistence diagram** is a Cartesian plot where each homological feature's lifetime interval is represented as a point $(\mathrm{birth}, \mathrm{death})$. By definition, they must lie at or above the line $y = x$.

We are mostly interested in the 1-dimensional homology, since the 0-dimensional homology is large identical to clustering analysis, and higher dimensional homologies are computationally expensive.

The subsets of data contributing to the top 10 most persistent features are then obtained with the help of Ripser, and transformed using **principal components analysis** (PCA) with 3 principal components. PCA is a dimensionality reduction technique using components that aim to capture the maximum amount of variance in the data.

Additionally, heatmaps of PCA loadings (how much each variable contributes to each component) are produced for features with interesting geometric shapes, such as a disk or a 3-dimensional loop. Variables without more than $|0.3|$ contribution to at least one principal component are omitted for readability.

Results and discussion

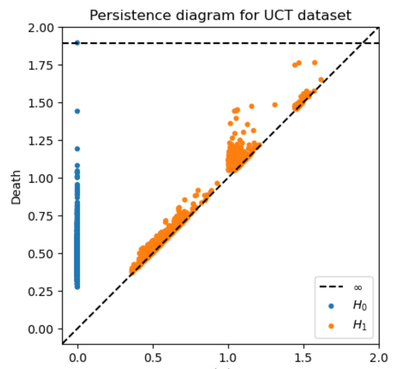
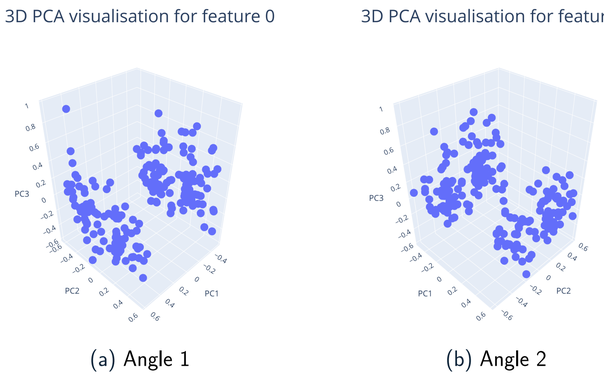


Figure 5. The persistence diagram for UCT dataset has some significant features

There are roughly 2 clusters that contain significant features (a cluster consists of features with similar birth times), possibly reflecting distinct underlying relations.

Some of the top 10 most persistent features are likely spurious since only a few data points contribute to them. Only the most illustrative examples are shown for brevity.



(a) Angle 1 (b) Angle 2

Figure 6. Feature 0 has two distinct clusters on the first component, and disks on the other two components

Results and discussion cont.

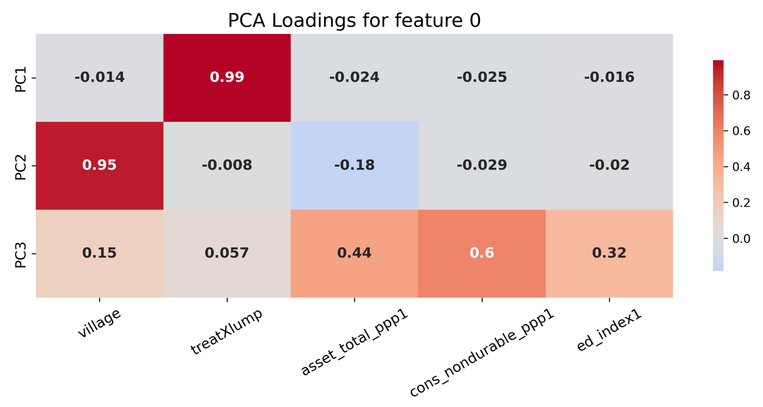
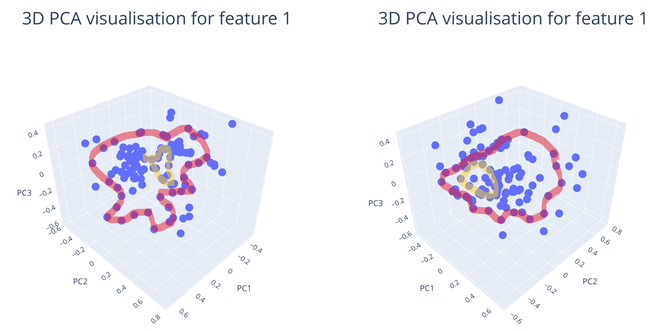


Figure 7. Loadings (contributions to component) for feature 0

PC1 almost exclusively represents the lump sum indicator, which separates households into two clusters. PC2 largely represents village number, a categorical variable, whereas PC3 captures the amount of household assets, consumption of nondurable goods and education index. The annuli formed on PC2 and PC3 highlight significant intra-village differences in these variables. The 1-dimensional holes, positioned near the average for PC2 and PC3, likely represent a lack of households that are perfectly average in these variables, and/or a lack of households in villages numbered near the middle.



(a) Angle 1 (b) Angle 2

Figure 8. As shown by the loops and inner clusters, feature 1's component 3 is negatively correlated with component 2. Rough shapes of the loops are highlighted

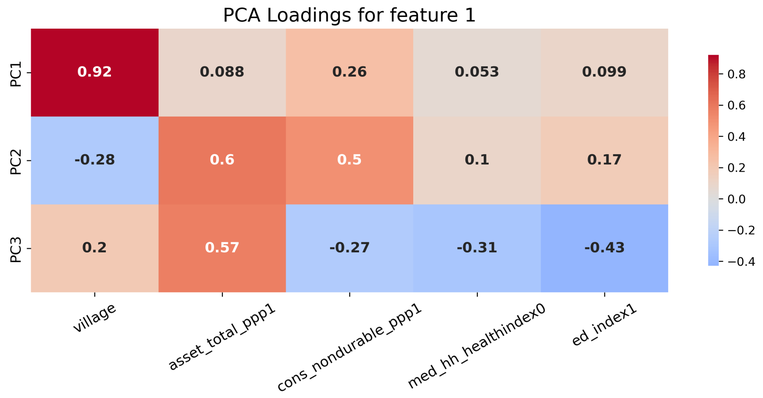


Figure 9. Loadings (contributions to component) for feature 1

The loadings for PC2 and PC3 indicate that the consumption of nondurable goods is somewhat negatively correlated with education index and health index. This could correspond to the idea that households in worse conditions are more likely to spend the cash on (nondurable) necessities in order to survive. Meanwhile, the amount of household assets appears uninformative.

Conclusion

- We have found interesting topological structures (disks, loops) in the dataset using 1-dimensional persistent homology, and made relevant visualisations and real-world interpretations of them with the help of PCA.
- Higher-dimensional homologies were not computed due to time constraints; for a dimension k, Ripser runs in $O(n^3)$ in the total number of $k+1$, k and $(k-1)$-simplices. [3]
- Further research is needed to investigate different types of datasets, and to explore other techniques in TDA.

Acknowledgements

I would like to extend my gratitude to my supervisors, Dr. Paul Fijn and Dr. TriThang Tran, for their guidance and support throughout this program.

References

[1] Ulrich Bauer.
Ripser: efficient computation of Vietoris-Rips persistence barcodes.
J. Appl. Comput. Topol., 5(3):391–423, 2021.

[2] Johannes Haushofer and Jeremy Shapiro.
The short-term impact of unconditional cash transfers to the poor: experimental evidence from kenya.
The Quarterly Journal of Economics, 131(4):1973–2042, 2016.

[3] Raúl Rabadán and Andrew J Blumberg.
Topological data analysis for genomics and evolution: topology in biology.
Cambridge University Press, 2019.

siyuanzhang@student.unimelb.edu.au