Comparing Distance Metrics on Vectorized Persistence Summaries

Brittany Terese Fasy(1) , Yu Qin(2) , Brian Summa(2) , Carola Wenk(2)

(1)Montana State University (2)Tulane University

Overview

Background:

- Topological data analysis (TDA) is powerful due to the ability to capture shapes and structure in data[1].
- The persistence diagram (PD) is an important tool in TDA for encoding an abstract representation of the homology of a shape at different scales.

Motivation:

- PD needs to be vectorized in order to apply it in machine learning and statistical tasks.
- The distance metric relationship between the vectorized persistence summaries and the original PDs has not been studied before.

Our Contributions:

- Studying the correlation between the distance of vectorized persistence summaries and distance of original (non-vectorized) PDs.
- Providing a new perspective on the use of TDA in machine learning.

I Topological Descriptors

Persistence diagrams and metrics

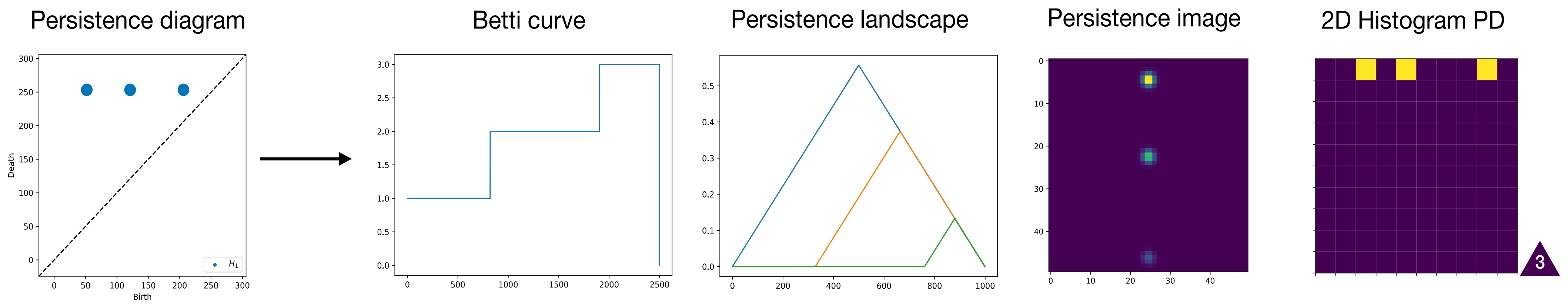
Given two persistence diagrams D_1 and D_2, the common distance metric is p–Wasserstein distance. In this work we use the $p = 1, \infty$,, both distances can be defined by finding an optimal matching $M \subset D_1 \times D_2$, considering both matched points $(x, y) \in M$ as well as unmatched points $M^c \subseteq D_1 \cup D_2$. For the p–Wasserstein distance between D_1 and D_2 is defined as :

$$W_{p(D_1,D_2)} := \inf_{M} \left(\sum_{(x,y)\in M} \|x-y\|_\infty^p + \sum_{x\in M^c} |x_1 - x_2|^p \right)^{\frac{1}{p}}$$

where $\|\cdot\|_\infty$ denotes the ∞-norm over the extended plane, and M ranges over all matchings between D_1 and D_2[2].

Read the paper

A persistence diagram and its four vectorized summaries



II Evaluation Method

Let Q be a given set of persistence diagrams, and let S be a function mapping each $q \in Q$ to a vectorized summary of q (e.g., a Betti Curve). Then for any query $q \in Q$ we calculate the distance to each other $p \in Q$, $p \neq q$, using the vectorized summary's metric $d_s(S(q), S(p))$.

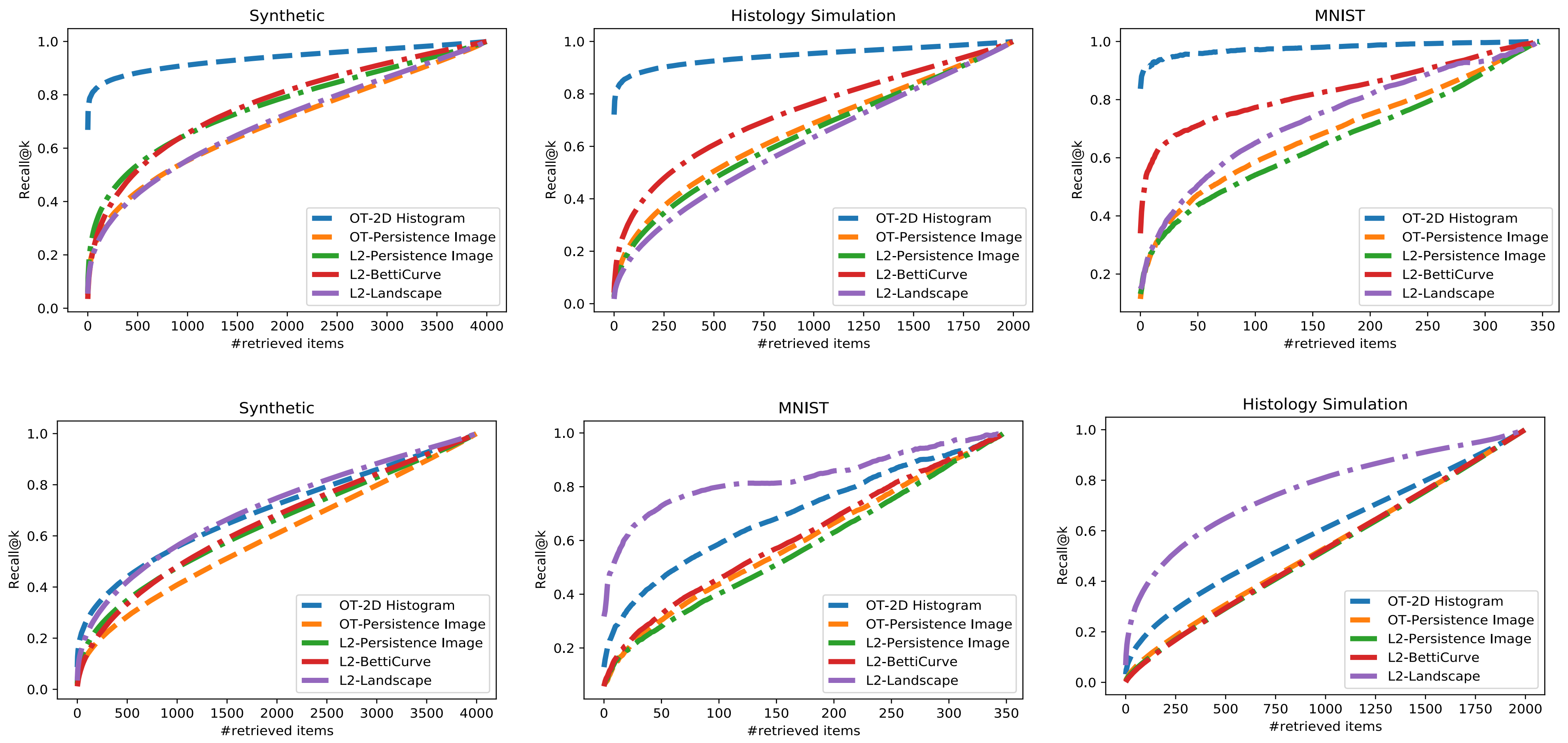
For any $k \in [1, |Q|]$ let $R_{d_s}(q, k) \subseteq Q$ consist of the k-nearest neighbors to q using $d_s(S(\cdot), S(\cdot))$. We compare this set to the set $R_{W_p}(q, k) \subseteq Q$ which consists of all k -nearest neighbors to k in $W_p(\cdot)$, here $p = 1, \infty$; using the original persistence diagrams (not summaries). We then evaluate the quality of d_s using the Recall@k:

$$Recall@k = \frac{1}{k|Q|} \sum_{q\in Q} \left| R_{d_s}(q,k) \cap R_w(q,k) \right|$$

We chose this particular evaluation measure because it can directly reflect the k-nearest neighbor consistency of the vectorized persistence summaries.

III Results

Recall rate performance on three datasets (first row $p = 1$, second row $p = \infty$)



Each curve indicates one vectorized persistence summary, and the first part of the label is the abbreviation of distance metric used in that approach. Here the x-axis refers to the range of datasets, and the value of the y-axis aggregated by the mean of overall k, where k equals the x-axis(#retrieved items). In the first row, the 2D Histogram in OT metric has the best performance for three datasets. In the second row, the persistence landscape in L2 outperforms other summaries.

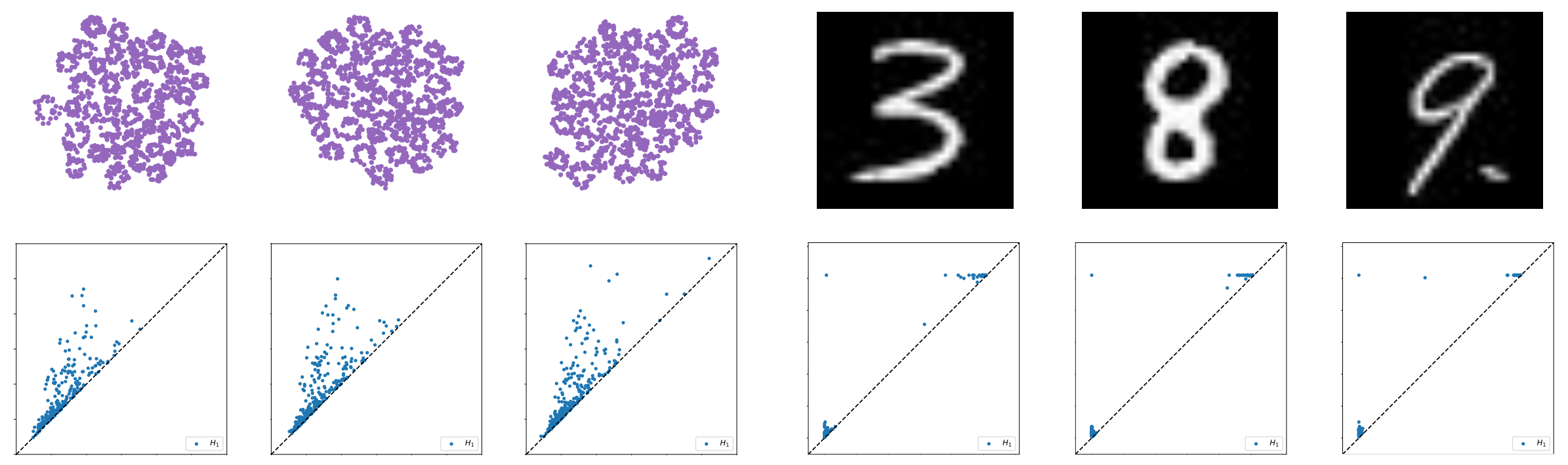
Vectorized summaries of persistence diagrams

Data in a PD is not amenable to many tasks. One way to resolve this issue is to transform a PD into a vectorized summary, which can be easily used in machine learning tasks. In this paper, we study the following four vectorizations that can be used to summarize a PD.

- Betti Curve
- Persistence Landscape
- Persistence Image
- 2D Histogram

As these summaries are vectors, we can compare two summaries using the L_p-distance and optimal transport (OT) metric.

Examples and corresponding persistence diagrams



Left: examples for Histology Simulation[3](top) and their PD via Rips filtration (bottom). Right: examples for MNIST (top) and their PD via sublevel filtration (bottom).

IV Discussion

- This study is the first work for metrics on vectorized persistence summaries.
- It is the first step toward enhancing our understanding of persistence diagram comparison.
- Our work helps to boost the usage of TDA in machine learning by determining the vectorized persistence summaries that can be better represented original persistence diagram.

Main References

[1] RIECK, B., SADLO, F., AND LEITTE, H. Topological machine learning with persistence indicator functions. arXiv preprint arXiv:1907.13496 (2019).

[2] COHEN-STEINER, D., EDELSBRUNNER, H., AND HARER, J. Stability of persistence diagrams. Discrete & Computational Geometry 37, 1 (2007), 103–120.

[3] FASY, B. T., PAYNE, S., SCHENFISCH, A., SCHUPBACH, J., AND STOUFFER, N. Simulating prostate cancer slide scans. In preparation, 2020.