Wasserstein Auto-Encoders of Merge Trees (and Persistence Diagrams)

M. Pont[1] J. Tierny[1]

[1] CNRS, Sorbonne Université (LIP6)

{mathieu.pont, julien.tierny}@sorbonne-universite.fr

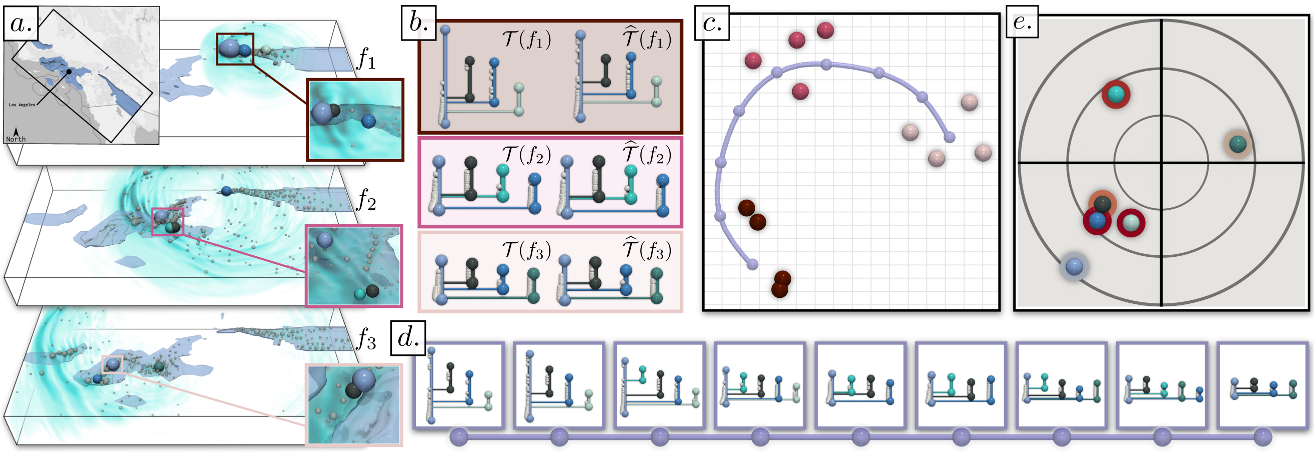

FIGURE 1 – *Visual analysis of the* Earthquake *ensemble* $\big((a)$ *each ground-truth class is represented by one of its members*$\big)$*, with our Wasserstein Auto-Encoder of Merge Trees (MT-WAE). We apply our contributions to merge tree compression* $\big((b)$*, right*$\big)$ *by simply storing their coordinates in the last decoding layer of our network. We exploit the latent space of our network to generate 2D layouts of the ensemble* (c)*. In contrast to classical auto-encoders, MT-WAE explicitly manipulates merge trees at each layer of the network, which results in improved accuracy and interpretability. Specifically, the reconstruction of user-defined locations* $\big((c)$*, purple*$\big)$ *enables an interactive exploration of the latent space : the reconstructed curve* (d) *enables a continuous navigation between the clusters* $\big($*from dark red to pink and light pink,* $(c)\big)$*. MT-WAE also supports persistence correlation views* (e) *(adapted from [1]), which reveal the barycenter's persistent features which exhibit the most variability in the ensemble (far from the center). Finally, by tracking the persistence evolution of individual features as they traverse the network down to its latent space, we introduce a* Feature Latent Importance *measure, which identifies the most* informative *features within the ensemble* $\big((e)$*, red circles*$\big)$*.*

Abstract

This paper presents a computational framework for the Wasserstein auto-encoding of merge trees (MT-WAE), a novel extension of the classical auto-encoder neural network architecture to the Wasserstein metric space of merge trees. In contrast to traditional auto-encoders which operate on vectorized data, our formulation explicitly manipulates merge trees on their associated metric space at each layer of the network, resulting in superior accuracy and interpretability. Our novel neural network approach can be interpreted as a non-linear generalization of previous linear attempts [1] at merge tree encoding. It also trivially extends to persistence diagrams. Extensive experiments on public ensembles demonstrate the efficiency of our algorithms, with MT-WAE computations in the orders of minutes on average. We show the utility of our contributions in two applications adapted from previous work on merge tree encoding [1]. First, we apply MT-WAE to merge tree compression*, by concisely representing them with their coordinates in the final layer of our auto-encoder. Second, we document an application to* dimensionality reduction*, by exploiting the latent space of our auto-encoder, for the visual analysis of ensemble data. We illustrate the versatility of our framework by introducing two penalty terms, to help preserve in the latent space both the Wasserstein distances between merge trees, as well as their clusters. In both applications, quantitative experiments assess the relevance of our framework. Finally, we provide a C++ implementation that can be used for reproducibility.*

Index Terms

Topological data analysis, ensemble data, merge trees, persistence diagrams.

1 Introduction

With the recent advances in the development of computation hardware and acquisition devices, datasets are constantly increasing in size. This size increase induces an increase in the geometrical complexity of the features present in the datasets, which challenges interactive data analysis and interpretation. To address this issue, Topological Data Analysis (TDA) [2] has shown over the years its ability to reveal, in a generic, robust and efficient manner, the main structural patterns hidden in complex datasets, in particular for visual data analysis tasks [3]. Among the representations studied in TDA, the merge tree [4] (Fig. 2) has been prominent in data visualization [5, 6, 7].

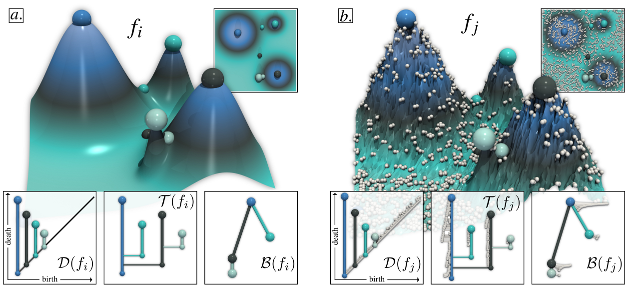

FIGURE 2 – *Illustration of the topological descriptors considered in this work, on a clean* (a) *and a noisy* (b) *variant of a 2D toy dataset. For all descriptors, the color code indicates the persistence of the corresponding saddle-maximum pair. Critical points are represented with spheres (larger ones for maxima). Persistence diagrams, merge trees and branch decomposition trees (BDTs) are respectively represented in the left, center and right insets. For both datasets, the four main features (the larger hills) are represented with salient pairs in the diagram and the merge tree. To avoid clutter in the visualization, the branches with low persistence (less than* 10% *of the function range) are rendered with small white arcs while larger, and colored arcs represent persistent branches (more than* 10% *of the function range). Figure adapted from [8, 1]*

In addition to the increase in geometrical complexity discussed above, a new challenge has recently emerged in many applications, with the notion of *ensemble datasets*. Such datasets encode a given phenomenon not only with a single dataset, but with a collection of datasets, called *ensemble members*. In that context, the topological analysis of an ensemble dataset consequently yields an ensemble of corresponding topological representations (e.g. one merge tree per ensemble member).

Then, developing statistical analysis tools to support the interactive analysis and interpretation of ensemble data becomes an important challenge. Recently, several works explored this direction, in particular with the notion of *average topological representation* [9, 10, 11, 12, 8]. These approaches can produce a topological representation which nicely summarizes the ensemble. Moreover, their application to clustering [8] reveal its main trends. However, they do not provide any hints regarding the variability of the features in the ensemble. For this, Pont et al. [1] recently extended the notion of principal geodesic analysis to ensembles of merge trees. However, this approach implicitly assumes a linear relation between the merge trees of the ensemble. Specifically, it assumes that merge tree branches evolve linearly (in the birth/death space) within the ensemble.

2 Contributions

This paper addresses this issue with a novel formulation based on neural networks and introduces the first framework for the non-linear encoding of merge trees, hence resulting in superior accuracy. Specifically, we formulate merge tree non-linear encoding as an auto-encoding problem. We contribute a novel neural network called *Wasserstein Auto-Encoder of Merge Trees*. This network is based on a novel layer model, capable of processing merge trees natively, without pre-vectorization. We believe this contribution to be of independent interest, as it enables an accurate and interpretable processing of merge trees by neural networks (without restrictions to auto-encoders). We contribute an algorithm for the optimization of such a network. Similarly to previous linear attempts [1], since our approach is based on the Wasserstein distance between merge trees [8], which generalizes the Wasserstein distance between persistence diagrams [2], our framework trivially extends to persistence diagrams by simply adjusting a single parameter.

3 Applications

We illustrate the relevance of our contributions for visual analysis with two applications. Firstly, data reduction, where we describe how to adapt previous work [1] to our novel non-linear framework, in merge tree compression applications. Specifically, the merge trees of the input ensemble are significantly compressed, by solely storing the final decoding layer of the network, as well as the coordinates of the input trees in this layer. We illustrate the interest of our approach with comparisons to linear encoding [1] in the context of feature tracking and ensemble clustering applications. Secondly, dimensionality reduction, for which we describe how to adapt previous work [1] to our novel non-linear framework. Specifically, each tree of the ensemble is embedded as a point in a planar view, based on its coordinates in our auto-encoder's latent space. To illustrate the versatility of our framework, we introduce two penalty terms, to improve the preservation of clusters and distances between merge trees.

4 Acknowledgments

This work is partially supported by the European Commission grant ERC-2019-COG "TORI" (ref. 863464, https ://erc-tori.github.io/). The code of this paper [13] is integrated in TTK (https ://topology-tool-kit.github.io/).

Appendix

Published in IEEE TVCG (Transactions on Visualization and Computer Graphics). DOI : 10.1109/TVCG.2023.3334755.

Références

[1] Mathieu Pont, Jules Vidal, et Julien Tierny. Principal Geodesic Analysis of Merge Trees (and Persistence Diagrams). *IEEE TVCG*, 29(2) :1573–1589, 2023.

[2] H. Edelsbrunner et J. Harer. *Computational Topology : An Introduction*. American Mathematical Society, 2009.

[3] C. Heine, H. Leitte, M. Hlawitschka, F. Iuricich, L. De Floriani, G. Scheuermann, H. Hagen, et C. Garth. A survey of topology-based methods in visualization. *CGF*, 35(3) :643–667, 2016.

[4] H. Carr, J. Snoeyink, et U. Axen. Computing contour trees in all dimensions. Dans *Symp. on Dis. Alg.*, 2000.

[5] Hamish A. Carr, Jack Snoeyink, et Michiel van de Panne. Simplifying Flexible Isosurfaces Using Local Geometric Measures. Dans *IEEE VIS*, 2004.

[6] P.T. Bremer, G. Weber, J. Tierny, V. Pascucci, M. Day, et J. Bell. Interactive exploration and analysis of large scale simulations using topology-based data segmentation. *IEEE TVCG*, 17(9) :1307–1324, 2011.

[7] Alexander Bock, Harish Doraiswamy, Adam Summers, et Cláudio T. Silva. TopoAngler : Interactive Topology-Based Extraction of Fishes. *IEEE TVCG (IEEE VIS)*, 24(1) :812–821, 2018.

[8] Mathieu Pont, Jules Vidal, Julie Delon, et Julien Tierny. Wasserstein Distances, Geodesics and Barycenters of Merge Trees. *IEEE TVCG (IEEE VIS)*, 28(1) :291–301, 2022.

[9] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, et John Harer. Fréchet Means for Distributions of Persistence Diagrams. *DCG*, 52(1) :44–70, 2014.

[10] Théo Lacombe, Marco Cuturi, et Steve Oudot. Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport. Dans *NIPS*, 2018.

[11] Jules Vidal, Joseph Budin, et Julien Tierny. Progressive Wasserstein Barycenters of Persistence Diagrams. *IEEE TVCG (IEEE VIS)*, 26(1) :151–161, 2020.

[12] Lin Yan, Yusu Wang, Elizabeth Munch, Ellen Gasparovic, et Bei Wang. A structural average of labeled merge trees for uncertainty visualization. *IEEE TVCG (IEEE VIS)*, 26(1) :832–842, 2019.

[13] Mathieu Pont et Julien Tierny. Wasserstein auto-encoders of merge trees (and persistence diagrams). *IEEE Transactions on Visualization and Computer Graphics*, pages 1–16, 2023.