Principal Geodesic Analysis of Merge Trees (and Persistence Diagrams)

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

[1] CNRS, Sorbonne Université (LIP6)

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

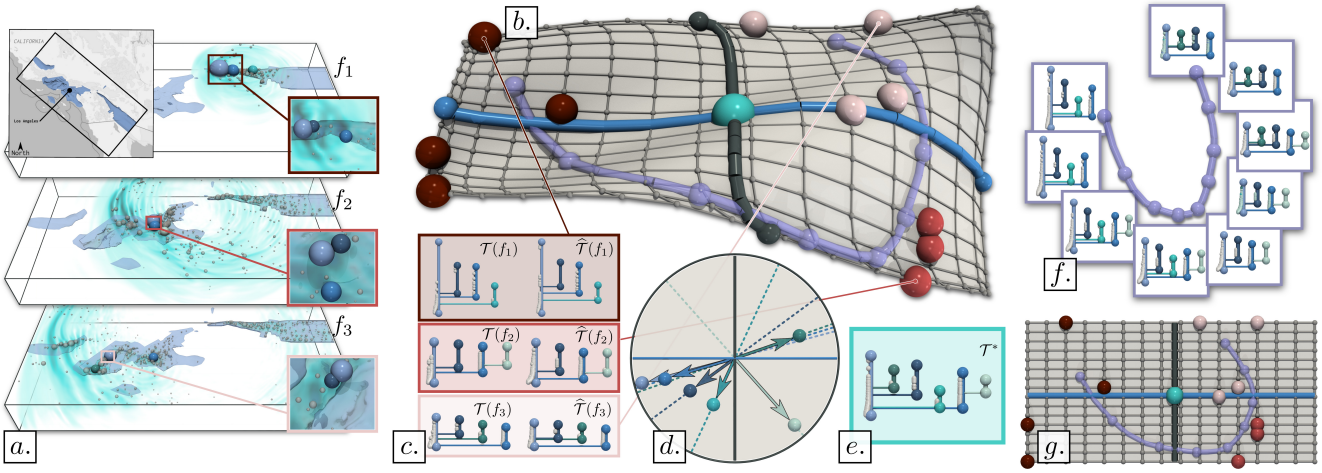

FIGURE 1 – *Visual analysis of the Earthquake ensemble with Merge Tree Principal Geodesic Analysis (MT-PGA, (a) : one member per ground-truth class). Our framework computes a coordinate system (b) for the Wasserstein metric space of merge trees B by adjusting geodesic axes (blue and black , (b)) to optimize a fitting energy. This enables the adaptation to merge trees of typical applications of Principal Component Analysis, such as data reduction, where the input trees are accurately reconstructed ((c), right), by simply storing their MT-PGA coordinates, or dimensionality reduction. MT-PGA enables the computation of a Principal Geodesic Surface (b), which complements its planar layout (g) by better conveying visually the curved nature of B. MT-PGA supports the efficient reconstruction of user-defined locations, for the interactive exploration of B : the reconstruction of the purple curve (f) enables the navigation from the trees of the first cluster (dark red, (b)) to the second (orange, (b)) and third (pink, (b)) clusters. MT-PGA also introduces Persistence Correlation Views (d) which enable the visual identification of the features which are the most responsible for the variability in the ensemble (high correlation, near the disk boundary, (d)) as well as their direct inspection in the data (matching colors (a)).*

Abstract

This paper presents a computational framework for the Principal Geodesic Analysis of merge trees (MT-PGA), a novel adaptation of the celebrated Principal Component Analysis (PCA) framework [1] to the Wasserstein metric space of merge trees [2]. We formulate MT-PGA computation as a constrained optimization problem, aiming at adjusting a basis of orthogonal geodesic axes, while minimizing a fitting energy. We introduce an efficient, iterative algorithm which exploits shared-memory parallelism, as well as an analytic expression of the fitting energy gradient, to ensure fast iterations. Our approach also trivially extends to extremum persistence diagrams. Extensive experiments on public ensembles demonstrate the efficiency of our approach – with MT-PGA computations in the orders of minutes for the largest examples. We show the utility of our contributions by extending to merge trees two typical PCA applications. First, we apply MT-PGA to data reduction *and reliably compress merge trees by concisely representing them by their first coordinates in the MT-PGA basis. Second, we present a* dimensionality reduction *framework exploiting the first two directions of the MT-PGA basis to generate two-dimensional layouts of the ensemble. We augment these layouts with persistence correlation views, enabling global and local visual inspections of the feature variability in the ensemble. In both applications, quantitative experiments assess the relevance of our framework. Finally, we provide a C++ implementation that can be used to reproduce our results.*

Index Terms

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

1 Introduction

Wether they are acquired or simulated, modern datasets are constantly gaining in detail and complexity, given the continuous improvement of acquisition devices or computing resources. This geometrical complexity is a difficulty for interactive data analysis and interpretation. This observation motivates the development of concise yet informative data representations, capable of encoding the main features of interest and visually representing them to the users. In that regard, Topological Data Analysis (TDA) [3] has demonstrated its ability to generically, robustly and efficiently reveal implicit structural patterns hidden in complex datasets.

Among the feature representations studied in TDA, the merge tree [4], which describes the global structure of the connected components of the sub-level sets of scalar datasets (Fig. 2), is a popular instance in the visualization community [5, 6, 7].

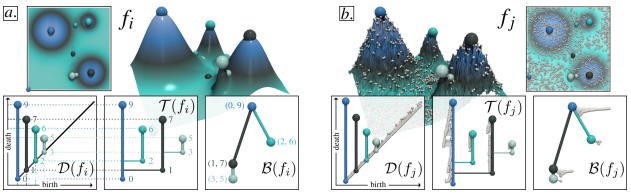

FIGURE 2 – *Critical points (spheres, larger radius : maxima), persistence diagram (left inset), merge tree (center inset) and branch decomposition tree (right inset) of a clean (a) and noisy (b) scalar field. In both cases, four main hills are clearly represented with salient features in the persistence diagram and the merge tree. Branches with low persistence (less than 10% of the function range) are shown with small white arcs.*

In many applications, on top of the increasing geometrical data complexity, an additional challenge emerges, related to *ensemble datasets*. These describe a phenomenon not only with a single dataset, but with a collection of datasets, called *ensemble members*, in order to characterize the variability of the phenomenon under study.

In principle, a topological representation (like the merge tree) can be computed for each ensemble member. While this strategy has several practical advantages (direct representations of the features of interest, reduced memory footprint), it shifts the analysis problem from an ensemble of datasets to an ensemble of merge trees. Then, a major challenge consists in designing statistical tools for such an ensemble of topological descriptors, to support its interactive analysis and interpretation. In this direction, a series of recent works focused on the notion of *average topological descriptor* [8, 9, 10, 11, 2], with applications to ensemble summarization and clustering. However, while such averages synthesize a topological descriptor which is well *representative* of the ensemble, they do not describe the topological variability of the ensemble.

2 Contributions

This paper addresses this issue and goes beyond simple averages by adapting the celebrated framework of Principal Component Analysis (PCA) [1] to ensembles of merge trees. For that, we introduce the novel notion of *"Merge-Tree Principal Geodesic Analysis"* (MT-PGA), which captures the most informative geodesics (i.e. analogs of straight lines on the abstract space of merge trees) given the input ensemble, hence facilitating variability analysis and visualization.

In particular, we formalize the computation of an orthogonal basis of principal geodesics in the Wasserstein metric space of merge trees [2] as a constrained optimization problem, inspired by previous work on the optimal transport of histograms [12, 13], which we extend and specialize to merge trees. We introduce an efficient iterative algorithm, which exploits an analytic expression of the energy gradient to ensure fast iterations.

Moreover, we document accelerations with shared-memory parallelism. Extensive experiments indicate that our algorithm produces bases of acceptable reconstruction quality within minutes, for real-life ensembles extracted from public benchmarks. Since our framework is based on the Wasserstein distance between merge trees [2], which generalizes the Wasserstein distance between persistence diagrams [8], it trivially extends to persistence diagrams by simply adjusting a parameter.

3 Applications

We illustrate the utility of our contribution in two applications. First, we show that the principal geodesic bases computed by our algorithm can result in an important compression of ensembles of merge trees, while still enabling a successful post-processing for typical visualization tasks such as feature tracking or ensemble clustering. Second, we present an extended application of our work to dimensionality reduction, for the visual inspection of the ensemble variability via two-dimensional embeddings, where we show that the views generated by our approach (e.g. the Principal Geodesic Surface and the Persistence Correlation View) preserve well the intrinsic metric between merge trees, as well as the global structure of the input ensembles, while enabling the visual inspection of the individual features which are the most responsible for the variability in the ensemble.

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 [14] is integrated in TTK (https ://topology-tool-kit.github.io/).

Appendix

Published in IEEE TVCG (Transactions on Visualization and Computer Graphics), Volume : 29, Issue : 2, 01 February 2023. DOI : 10.1109/TVCG.2022.3215001.

Références

[1] K Pearson. On Lines and Plances of Closest Fit to Sytems of Points in Space. *Philosophical Magazine*, 2 :559–572, 1901.

[2] Mathieu Pont, Jules Vidal, Julie Delon, et Julien Tierny. Wasserstein Distances, Geodesics and Barycenters of Merge Trees. *IEEE Transactions on Visualization and Computer Graphics*, 28(1) :291–301, 2022.

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

[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 Transactions on Visualization and Computer Graphics*, 17(9) :1307–1324, 2011.

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

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

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

[10] Jules Vidal, Joseph Budin, et Julien Tierny. Progressive Wasserstein Barycenters of Persistence Diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 26(1) :151–161, 2020.

[11] Lin Yan, Yusu Wang, Elizabeth Munch, Ellen Gasparovic, et Bei Wang. A structural average of labeled merge trees for uncertainty visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(1) :832–842, 2019.

[12] Vivien Seguy et Marco Cuturi. Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric. Dans *NeurIPS*, 2015.

[13] Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, et Nicolas Papadakis. Geodesic PCA versus Log-PCA of Histograms in the Wasserstein Space. *SIAM J. Sci. Comput.*, 40(2), 2018.

[14] Mathieu Pont, Jules Vidal, et Julien Tierny. Principal Geodesic Analysis of Merge Trees (and Persistence Diagrams). *IEEE Transactions on Visualization and Computer Graphics*, 29(2) :1573–1589, 2023.