On the Implementation of Geodesic Metric Spaces

Anna Calissano, Luís F Pereira, Jonas Lueg, Nina Miolane

On the Implementation of Geodesic Metric Spaces

Anna Calissano a.calissano@imperial.ac.uk
Inria Center at Université Côte d'Azur, Valbonne, France
now at Imperial College London, London, United Kingdom

Luís F. Pereira luis.gomes-pereira@inria.fr
Inria Center at Université Côte d'Azur, Valbonne, France

Jonas Lueg jonas.lueg@posteo.de
Georg-August-Universität, Göttingen, Germany

Nina Miolane ninamiolane@ucsb.edu
University of Santa Barbara, Santa Barbara, California, USA

Editor:

Abstract

Analysis of non-Euclidean data such as graphs and trees requires (specific) mathematical machinery due to their less-rich structure when compared to Euclidean spaces or smooth Riemannian manifolds. These spaces can still leverage the rich structure of the latter. For example, graph space results from quotienting out matrices endowed with the Frobenius metric by the permutation group, Billera–Holmes–Vogtmann (BHV) space strata are Euclidean, and wald space is embedded in the space of symmetric positive definite (SPD) matrices. We present a Python package for the analysis of data living in geodesic metric spaces – topological spaces equipped with a metric and a geodesic function where the metric is the length of the shortest geodesic joining two points. We describe the package structure, based on a point, a point set, and a metric built using geodesic metric space theory, and we provide three implementation examples. The package is implemented as a plug-in of the Geomstats Python package, allowing users to access and adapt the available geometrical and data analysis tools for strongly non-Euclidean data in a theoretically consistent way. The code is unit-tested and documented.

Keywords: geodesic metric spaces; BHV space; tree-valued data; graph-valued data; geometric data analysis.

1. Introduction

The analysis of non-Euclidean data such as trees and graphs has gathered scientific attention in the last years (Marron and Dryden, 2021; Calissano et al., 2024a; Billera et al., 2001; Garba et al., 2021; Severn et al., 2021; Huckemann and Eltzner, 2021). These data types are relevant to several applications: for example brain connectivity (Simpson et al., 2013; Durante et al., 2017; Ginestet et al., 2017; Calissano et al., 2024b), mobility (von Ferber et al., 2009), airlines trees (Feragen et al., 2013), and phylogenetics (Billera et al., 2001; Garba et al., 2021; Lueg et al., 2024). These data types are referred to as strongly non-Euclidean or beyond manifold data types (Marron and Dryden, 2021). To analyse such data, the majority of the methods relies on the theory of geodesic metric spaces, as the smooth or Riemannian manifold tools are not or only locally available. Geodesic metric

spaces are spaces with length minimizing geodesics well defined everywhere and a metric - i.e., a distance function - obtained by computing the length of such geodesics (Bridson and Haefliger, 1999). Common examples of geodesic metric spaces are non-manifold quotient spaces (i.e., quotient spaces without a well-defined tangent space), and stratified spaces (i.e., union of compatibly glued manifolds of potentially different dimensions, cf. Mather (1973)). Geodesic metric spaces have been used as a geometric embedding for the analysis of matrices, graphs, and trees (see Feragen and Nye (2020) for an overview). Within such spaces, a plethora of methods have been developed: for example principal component analysis (Wang and Marron, 2007; Feragen et al., 2011), and regression methods (Calissano et al., 2022; Severn et al., 2021; Petersen and Müller, 2019).

While many different implementations have been proposed in the context of manifolds, the implementation of geometrical frameworks and learning methods for geodesic metric spaces is scattered and undeveloped compared to the theory. Even if there exist many different packages for the analysis and the visualization of a single non-Euclidean datum such as a graph or a tree (EasyGraph (Aldridge, 2023), Networkx (Hagberg and Conway, 2020), iGraph (Csardi et al., 2006), nograph (Helmut, 2023), treelib (Xiaming, 2023), rustworkx (Treinish et al., 2021), Scikit-network (Bonald et al., 2020)), none of these packages offer a framework for the analysis of sets of such non-Euclidean data. To the best of our knowledge, the only package implementing methods for sets of graphs is Graspy (Chung et al., 2019), offering different alignment procedure for graphs.

We propose a module providing flexible abstract classes for the implementation of existing and novel geodesic metric spaces – including embeddings, stratified spaces, and quotient spaces. Such implementation can then be used within geometrical and learning methods requiring the notion of distance and/or geodesics. The module is implemented within Geomstats Python package and takes advantage - when possible - of the available manifold and statistical methods, i.e. both geometry and learning.

2. Theoretical Background: Geodesic Metric Spaces

Let $(\mathcal{X}, d)$ be a metric space, containing data objects $x \in \mathcal{X}$, where x might not have a trivial representation as a vector or a matrix. Each metric space induces an intrinsic metric d^* (or length metric), where the induced intrinsic distance $d^*(x, y)$ for two points $x, y \in \mathcal{X}$ is the infimum over the length of all continuous paths connecting x and y (cf. Bridson and Haefliger (1999)); and if $d = d^*$, then $(\mathcal{X}, d)$ is called a length metric space. Given two points $x, y \in \mathcal{X}$, a minimizing geodesic is a map $c : [0, l] \subset \mathbb{R} \to \mathcal{X}$ such that $x = c(0), \quad y = c(l)$ and $d(c(t'), c(t)) = |t' - t|$ for all $t, t' \in [0, l]$. A metric space $(\mathcal{X}, d)$ is called **geodesic metric space** if there every two points are joined by a geodesic; and one can show that each geodesic metric space is a length metric space. In contrast to the manifold setting, the definition of geodesic is not based on linear local approximation of the space via the tangent space, making geodesic metric spaces more general. A length metric space is often required for data analysis in metric spaces (e.g. computation of Fréchet means (Sturm, 2023)). In Figure 1, we give a visual example of the sphere (left), where the length metric is the intrinsic metric induced by the ambient Euclidean distance, and the tripod (right), a simple example of stratified space, where the metric and geodesics are defined through the union of segments on the strata (i.e., the rays).

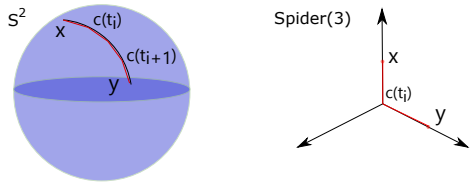

Figure 1: Length metric on a sphere and on the spider.

3. Package structure

We propose an abstract framework for the implementation of geodesic metric spaces to handle non-Euclidean data types (see Figure 2 for details). The structure consists of the following elements: (1) x: a data point representation, (2) $\mathcal{X}$: a point set, (3) $\gamma : \mathcal{X} \to \mathbb{R}$: a geodesic function - by default parametrized between $[0,1]$, (3) $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$: a point set metric, by default obtained by computing the length of the shortest geodesic. We describe and implement three examples taking full advantage of the new structure and the existing objects in Geomstats and thereby cover three types of mathematical structures in the literature: stratified spaces, quotient spaces, and general embeddings.

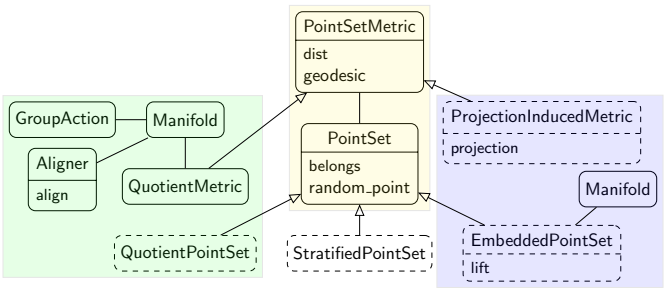

Figure 2: Implementation of geodesic metric spaces in Geomstats. (open triangle: inheritance)

Stratified spaces are spaces obtained by gluing together smooth manifolds of varying dimension (see Pflaum (2001) for a formal definition). Stratified spaces appear in many different applications: trees (Billera et al., 2001), graphs (Kolaczyk et al., 2020b), persistent diagrams (Turner et al., 2014), positive semi-definite matrices (Thanwerdas and Pennec, 2021; Takatsu, 2011).

Example 1 *BHV Space (Billera et al., 2001) is the seminal embedding for tree data. Consider the BHV space for rooted trees with n leaves: (1) x: the tree representation by splits and internal edge lengths;(2) $\mathcal{X}$: the point set consisting of* $(2n-3)!!$ *copied of Euclidean orthants of dimension* $n-2$*; (3)* $\gamma : \mathcal{X} \to \mathbb{R}$ *the geodesic function computed as in Owen and Provan (2010); (4)* $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *the point set metric obtained by computing the length of the shortest geodesic.*

Quotient spaces can be obtained by applying a group action G to a smooth manifold $\mathcal{M}$, often referred to as orbit spaces (Lee et al., 2017). If the action is not proper or free the resulting space is not a manifold, but it is a geodesic metric space. Examples are

graph space (Calissano et al., 2024a; Jain and Obermayer, 2009) and shape spaces (Kendall, 1984).

Example 2 *Graph Space is a space introduced to study node unlabeled graphs (Calissano et al., 2024a; Kolaczyk et al., 2020a; Jain and Obermayer, 2009). Every weighted directed graph G is represented as a weighted adjacency matrix* $\mathbb{R}^{n\times n}$*. Then, a permutation action is applied to the nodes of the graph, resulting in a natural embedding for graphs with no node labels. It is a discrete quotient space, but it is not a manifold. The child structure is: (1) x: an adjacency matrix representing a graph with n nodes; (2)* $\mathcal{X} = \mathbb{R}^{n\times n}/\mathbb{P}$*, where* $\mathbb{R}^{n\times n}$ *is the total space and* $\mathbb{P}$ *is the set of permutation matrices acting by conjugation. Every point in the space is an equivalence class* $[x] = \{pxp^t : p \in \mathbb{P}\}$*; (3)* $\gamma : \mathcal{X} \to \mathbb{R}$ *the geodesic function computed using an aligner map between* $a([x],[y]) = (x,y)$*. The aligner spans the discrete equivalence classes and returns optimally aligned adjacency matrices based on total space metric; (4)* $d_{\mathcal{X}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ *point set metric computed between aligned graphs and corresponding to length minimizing geodesics.*

A topological space $\mathcal{X}$ which can be embedded into an ambient space $\mathcal{M}$ via an injective continuous map (an embedding) $\mu\colon \mathcal{X} \hookrightarrow \mathcal{M}$ is called an embedded space. A simple example is the sphere S^2 embedded in $\mathbb{R}^3$. The implementation of these spaces takes particular advantage of selecting $\mathcal{M}$ among the available manifolds in Geomstats.

Example 3 *Wald space was introduced for the analysis of phylogenetic trees, embedding forests (i.e., trees allowing for isolated leaves) into the space of positive definite matrices of dimension n -* Sym_n^+ *(Garba et al., 2021; Lueg et al., 2024). The child structure is: (1) x a forest with n leaves; (2)* $\mathcal{X}$ *the set of forests; (3)* $\gamma : \mathcal{X} \to \mathbb{R}$ *the geodesic function computed using the embedding* $\mu : \mathcal{X} \to Sym_n^+$*. The ambient space* Sym_n^+ *can be equipped with different metrics and geodesics available in Geomstats; (4)* $d_{\mathcal{X}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ *point set metric computed using the map and the ambient space metric.*

4. Example of usage

As an example of implementation, we select the BHV space (Billera et al., 2001), as the most well known metric space for set of phylogenetic trees.

```
# create points
e1 = Split((0, 3, 4), (1, 2))
e2 = Split((0, 4), (1, 2, 3))
e3 = Split((0, 1, 4), (2, 3))
point_1 = Tree(splits=[e1, e2], lengths=gs.array([1., 0.5]))
point_2 = Tree(splits=[e1, e3], lengths=gs.array([2., 1.5]))
# instantiate the space
space = TreeSpace(n_labels=5, equip=False)
space.equip_with_metric(BHVMetric)
# compute distance
space.metric.dist(point_1, point_2)
# find geodesic
geodesic_func = space.metric.geodesic(point_2, point_3)
t = gs.array([0.2, 0.5])
geod_points = geodesic_func(t)
```

Acknowledgements

We would like to thank the whole Geomstats contributors and Xavier Pennec for general support, as well as Tom Nye for the insights on geodesic metric spaces geometry. The project was funded by the ERC Advanced grant 786854 on Geometric Statistics and by the Deutsche Forschungsgesellschaft via the Research Training Group 2088.

References

J. Aldridge. Easy graph 0.1, 2023. URL `https://pypi.org/project/easy_graph/`.

L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

T. Bonald, N. De Lara, Q. Lutz, and B. Charpentier. Scikit-network: Graph analysis in python. *The Journal of Machine Learning Research*, 21(1):7543–7548, 2020.

M. R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*. Springer, 1999.

A. Calissano, A. Feragen, and S. Vantini. Graph-valued regression: Prediction of unlabelled networks in a non-euclidean graph space. *Journal of Multivariate Analysis*, 190: 104950, 2022.

A. Calissano, A. Feragen, and S. Vantini. Populations of unlabelled networks: Graph space geometry and generalized geodesic principal components. *Biometrika*, 111(1):147–170, 2024a.

A. Calissano, T. Papadopoulo, X. Pennec, and S. Deslauriers-Gauthier. Graph alignment exploiting the spatial organization improves the similarity of brain networks. *Human Brain Mapping*, 45(1):e26554, 2024b.

J. Chung, B. D. Pedigo, E. W. Bridgeford, B. K. Varjavand, H. S. Helm, and J. T. Vogelstein. Graspy: Graph statistics in python. *J. Mach. Learn. Res.*, 20(158):1–7, 2019.

G. Csardi, T. Nepusz, et al. The igraph software package for complex network research. *InterJournal*, *complex systems*, 1695(5):1–9, 2006.

D. Durante, D. B. Dunson, and J. T. Vogelstein. Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112(520):1516–1530, 2017.

A. Feragen and T. M. W. Nye. 8 - statistics on stratified spaces. In X. Pennec, S. Sommer, and P. T. Fletcher, editors, *Riemannian Geometric Statistics in Medical Image Analysis*, pages 299 – 342. Academic Press, 2020. ISBN 978-0-12-814725-2. doi: https://doi.org/10.1016/B978-0-12-814725-2.00016-9.

A. Feragen, S. Hauberg, M. Nielsen, and F. Lauze. Means in spaces of tree-like shapes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 736–746, 2011.

A. Feragen, M. Owen, J. Petersen, M. Wille, L. Thomsen, A. Dirksen, and M. de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *Information Processing in Medical Imaging*, volume 7917 of *Lecture Notes in Computer Science*, pages 74–85. Springer, 2013.

M. K. Garba, T. M. Nye, J. Lueg, and S. F. Huckemann. Information geometry for phylogenetic trees. *Journal of Mathematical Biology*, 82(3):1–39, 2021.

C. E. Ginestet, J. Li, P. Balachandran, S. Rosenberg, and E. D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2): 725–750, June 2017.

A. Hagberg and D. Conway. Networkx: Network analysis with python. *URL: https://networkx. github. io*, 2020.

M. Helmut. nograph 3.1.0, 2023. URL `https://pypi.org/project/nograph/`.

S. F. Huckemann and B. Eltzner. Data analysis on nonstandard spaces. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(3):e1526, 2021.

B. J. Jain and K. Obermayer. Structure spaces. *Journal of Machine Learning Research*, 10 (Nov):2667–2714, 2009.

D. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

E. D. Kolaczyk, L. Lin, S. Rosenberg, J. Walters, and J. Xu. Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics*, 48(1):514–538, 2020a.

E. D. Kolaczyk, L. Lin, S. Rosenberg, J. Walters, J. Xu, et al. Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics*, 48(1):514–538, 2020b.

H. Lee, Z. Ma, Y. Wang, and M. K. Chung. Topological distances between networks and its application to brain imaging. *arXiv preprint arXiv:1701.04171*, 2017.

J. Lueg, M. Garba, T. Nye, and S. Huckemann. Foundations of the wald space for phylogenetic trees. *Journal of the London Mathematical Society*, 109(5):1–45, May 2024. ISSN 0024-6107. doi: 10.1112/jlms.12893.

J. S. Marron and I. L. Dryden. *Object oriented data analysis*. Chapman and Hall/CRC, 2021.

J. N. Mather. Stratifications and mappings. In *Dynamical systems*, pages 195–232. Elsevier, 1973.

M. Owen and J. S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):2–13, 2010.

A. Petersen and H.-G. Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.

M. Pflaum. *Analytic and geometric study of stratified spaces: contributions to analytic and geometric aspects*. Number 1768. Springer Science & Business Media, 2001.

K. E. Severn, I. L. Dryden, and S. P. Preston. Non-parametric regression for networks. *Stat*, 10(1):e373, 2021.

S. L. Simpson, R. G. Lyday, S. Hayasaka, A. P. Marsh, and P. J. Laurienti. A permutation testing framework to compare groups of brain networks. *Frontiers in computational neuroscience*, 7:171–184, 2013.

K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. 338, 2023. ISSN 9780821833834. doi: 10.1090/conm/338/06080.

A. Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48 (4):1005–1026, 2011.

Y. Thanwerdas and X. Pennec. Geodesics and curvature of the quotient-affine metrics on full-rank correlation matrices. In *Geometric Science of Information: 5th International Conference, GSI 2021, Paris, France, July 21–23, 2021, Proceedings 5*, pages 93–102. Springer, 2021.

M. Treinish, I. Carvalho, G. Tsilimigkounakis, and N. Sá. Rustworkx: A high-performance graph library for python. *arXiv preprint arXiv:2110.15221*, 2021.

K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distributions of persistence diagrams. *Disc. Comp. Geom.*, 52(1):44–70, 2014.

C. von Ferber, T. Holovatch, Y. Holovatch, and V. Palchykov. Public transport networks: empirical analysis and modeling. *The European Physical Journal B*, 68(2):261–275, 2009.

H. Wang and S. J. Marron. Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.

C. Xiaming. treelib 1.6.4, 2023. URL `https://pypi.org/project/treelib/`.