Manifold Modelling with Minimum Spanning Trees

Daniel Bot[1], Peiyang Huo[3], Alessio Arleo[2], Fernando Paulovich[2] & Jan Aerts[1,3]

[1]Data Science Institute, UHasselt, Belgium; [2] Dept of Mathematics & Computer Science, TU Eindhoven, Netherlands;
[3]Dept of Biosystems, KU Leuven, Belgium

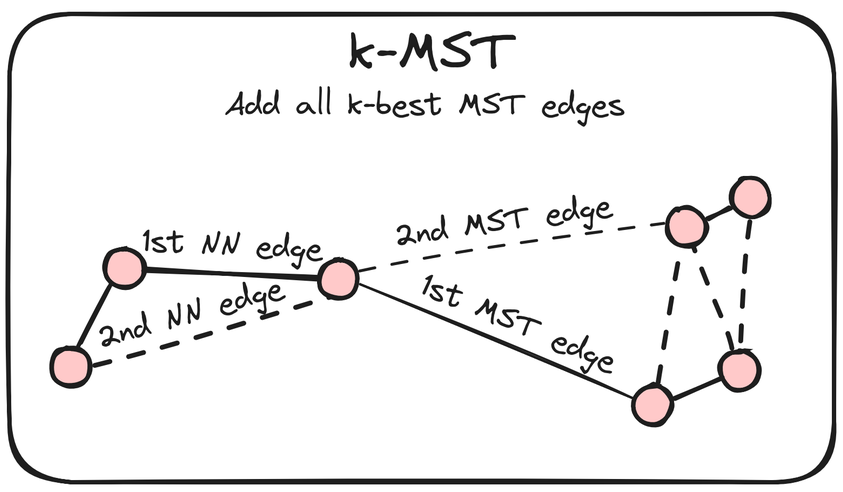
Contact: jan.aerts@kuleuven.be

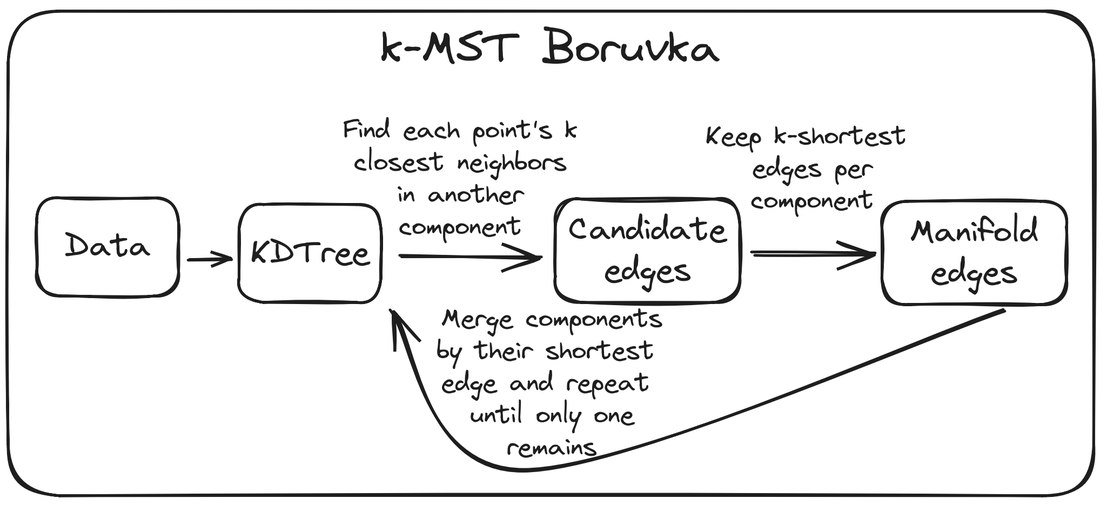
Background

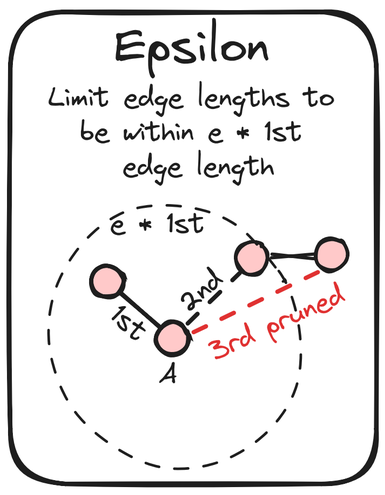
Recent dimensionality reduction algorithms assume data is uniformly sampled from an underlying manifold (e.g., [RS00, Ten00, vdMH08, MHM18]. While some algorithms are fairly robust against this assumption, there are scenarios in which more than the k-nearest neighbours (k-NN) are needed to approximate a manifold. This poster presents a k-nearest Minimum Spanning Tree (k-MST) manifold approximation approach that can deal with non-uniform sampling.

Methodology

The k-nearest Minimum Spanning Tree (k-MST) adapts MST unions from Pathfinder networks for distances. Our implementation adapts the Boruvka Algorithm from [MC23] to keep the k shortest edges between connected components. Since data points start as distinct components, all k-NN edges are included in the k-MST. A distance threshold ε provides an upper distance limit for the 2-to-k alternative edges to avoid shortcuts or sparsify the network.

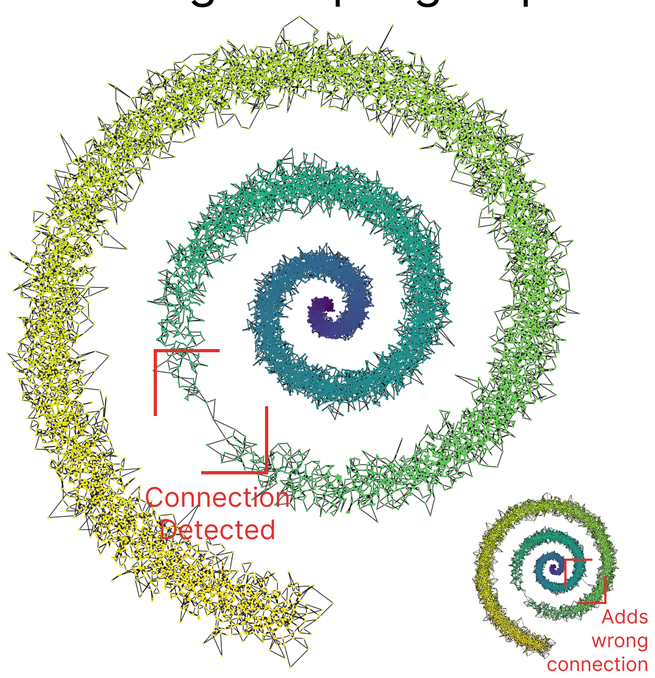






How do k-NNs and k-MSTs compare in modelling non-uniformly sampled manifolds?

Crossing Sampling Gaps

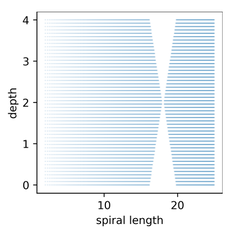


2-MST *time: 105ms*

5-NN *time: 124ms*

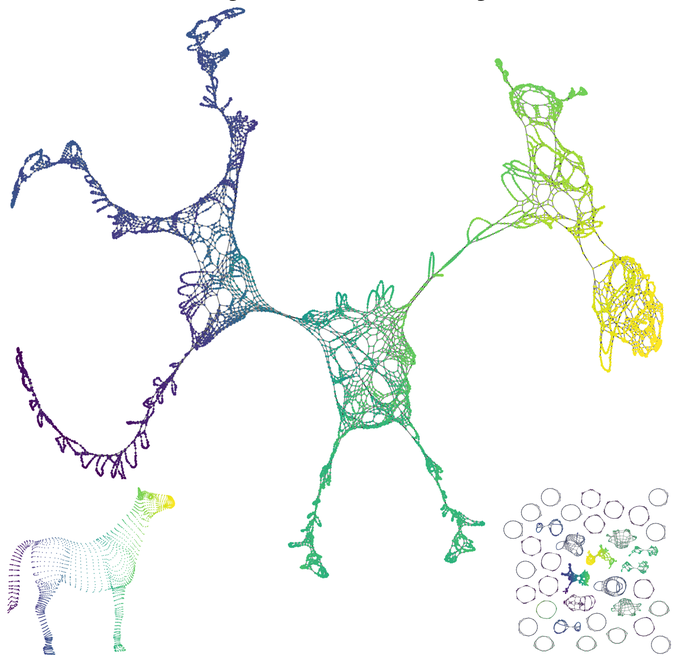
The ability to cross sampling-gaps was evaluated on a non-uniformly sampled Swiss Roll built from lengths `l` and depths `d`:

$$x = sl^2\cos(l) + \mathcal{N}(0, nl)$$
$$y = sl^2\sin(l) + \mathcal{N}(0, nl)$$
$$z = d + \mathcal{N}(0, nl)$$
$$s = 0.03, n = 0.0395$$



Some samples were removed along the manifold, resulting in a dataset with `22.196` samples and `3` features.

Layout Quality



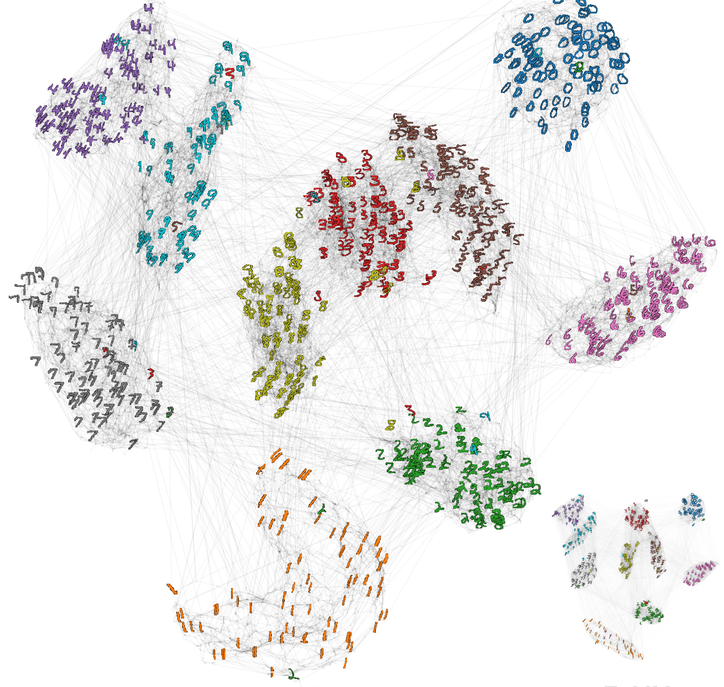
side-view

3-MST *time: 40ms sortedness: 0.90*

5-NN *time: 68ms sortedness: 0.72*

The methods' quality as dimensionality reduction approach was quantified by the Sortedness [PSNCP23] of their force-directed layouts [Hu05] on a horse-shaped dataset [SP04] with `8.431` samples and `3` features. The `3`-MST recovers a small, connected graph. The `5`-NN does not create a fully connected graph. At `10`-NN the Sortedness improves to `0.89`, but the graph remains disconnected.

Clustered Data



2-MST *time: 647ms edges: 201.032*

5-NN *time: 68ms edges: 427.046*

The methods' behavior on data with clusters was compared using MNIST [LBBH98], which has `70.000` samples and `784` features. UMAP layouts show the created networks. The `2`-MST was computed with `ε=1.1` to sparsify the graph. It balances local and global structures in a small model that is relatively cheap to lay out. The `5`-NN creates a larger network with more in-cluster connectivity.

Conclusions

k-MST discovers connectivity at all distance scales, balancing local and global structure and creating small models. A limitation of our implementation is the compute cost associated with KDTrees on some data sets. A NNdescent-based approximation for MSTs could make our techniques computationally competitive for larger datasets.

References

- [Hu05] HU Y. W. R. I.: Efficient and High Quality Force-Directed Graph Drawing. Math. J. 10, 1 (2005), 37–71.
- [MC23] MCINNES L., CONTRIBUTORS: Fast HDBSCAN (version 0.1.2), 2023. https://github.com/TutteInstitute/fast_hdbscan/release.
- [MHM18] MCINNES L., HEALY J., MELVILLE J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.
- [PSNCP23] PEREIRA-SANTOS D., NEVES T. T. A. T., CARVALHO A.C. P. L. F. D., PAULOVICH F. V.: Nonparametric Dimensionality Reduction Quality Assessment based on Sortedness of Unrestricted Neighborhood. In EuroVis Workshop on Visual Analytics (EuroVA) (2023), Angelini M., El-Assady M., (Eds.), The Eurographics Association.
- [RS00] ROWEIS S. T., SAUL L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science (80-.). 290, 5500 (dec 2000), 2323–2326.
- [SP04] SUMNER R. W., POPOVI ´C J.: Deformation transfer for triangle meshes. ACM Trans. Graph. 23, 3 (aug 2004), 399–40.
- [Ten00] TENENBAUM J. B.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science (80-.). 290, 5500 (dec 2000), 2319–2323.
- [vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (nov 2008), 2579–2625.
- [LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient based learning applied to document recognition. Proc. IEEE 86, 11 (1998), 2278–2324.
- [AKM17] ARLEO A., KWON O. H., MA K. L.: GraphRay: Distributed pathfinder network scaling. 2017 IEEE 7th Symp. Large Data Anal. Vis. LDAV 2017 2017-Decem (2017), 74–83.

This work was supported by Hasselt University BOF grant [BOF20OWB33] and KU Leuven grant ITP-E5160-STG/23/040.