Research Directions to Validate Topological Models of Multi-Dimensional Data

Nello Blaser
Department of Informatics
Center for Data Science
University of Bergen
Norway
nello.blaser@uib.no

Michaël Aupetit
Qatar Computing Research Institute
Hamad Bin Khalifa University
Qatar
maupetit@hbku.edu.qa

Abstract

Various topological models of multi-dimensional data have been proposed for different applications. One of the main issues is to evaluate how correct these models are given the stochastic nature of the data source typical of exploratory data analysis and machine learning settings. We propose research directions to validate the quality of the Mapper and the Generative Simplicial Complex, two models that compute simplicial complexes from multi-dimensional data.

1 Context

Topological methods in machine learning aim to quantitatively encode shape information from data points. This includes multiscale summaries (e.g. persistent homology [1]), topology-preserving dimensionality reduction (e.g. UMAP [2][3]), or encoding of topological information in simplicial complexes (e.g. Generative Simplicial Complexes [4, 5], Mapper [6]). Validating simplicial complex encodings of high-dimensional data are the focus of this manuscript.

Model validation consists of selection and evaluation, two core concepts in machine learning. Model selection aims at finding the best (usually simplest) of otherwise similarly accurate models, for example by tuning hyperparameters to change its topology (number of hidden layers in a neural network, number of components in a mixture model). Model evaluation aims at estimating how well a model generalizes to unseen data. Both model selection and evaluation rely strongly on a measure that allows to compare different models, which we call a validation measure. However, there is a clear lack of standard model validation measures for topological methods.

For supervised learning, model selection and evaluation principles are well established for instance using cross-validation and prediction loss minimization. For unsupervised learning like clustering, there are often specialized model evaluation metrics. Clustering can be seen as a special case of finding connected components of simplicial complex encodings [3]. Clusters are typically evaluated using either external validation such as class labels, or internal validation combining compactness and separation measures [7]. External validation is also possible and commonly used for simplicial complex encodings [8, 9]. However, we are not aware of any intrinsic validation methods, which would focus on topological properties of the simplicial complex encodings in relation to statistical properties of the data.

2 Encoding topology from multi-dimensional data

We propose to consider the Mapper and the Generative Simplicial Complex, two different models that build simplicial complexes from multidimensional data. We want to understand how their parameters

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

affect the connection between the geometry and the probability distribution of the data, and the topology of the simplicial complex encoding them. Both Mapper and Generative Simplicial Complex assume data is generated from a latent manifold $M \subset \mathbb{X} = \mathbb{R}^d$ corrupted by some additive noise with density g and aim to infer properties of the topology of M from the data sample X.

The **Mapper** algorithm approximates the Reeb graph of the manifold M, equipped with a continuous filter function f from $\mathbb{R}^d$ to $\mathbb{R}$. In particular, it studies the connected components of the level sets of the filter function [6]. The Mapper algorithm divides $\mathbb{R}$ into overlapping intervals, clusters the data in preimages of these intervals and computes the nerves of the resulting cover of X.

Hyperparameters that have to be selected are the interval sizes, overlap sizes, clustering method, and hyperparemeters associated with the chosen clustering method. In addition, the filter function f is often chosen from data, complicating model comparison further. In applications, hyperparameters are often tuned manually and lack theoretic grounding, although there exists some a priori methods for hyperparameter selection [10].

The **Generative Simplicial Complex** (GSC) [4, 5] extends Gaussian Mixture Models (GMM) [11] to higher dimensional simplicial complexes. The GSC is based on a set of K landmark vectors $W \in \mathbb{X}^K$ (*e.g.* the component means of a GMM) and a simplical complex S of W (*e.g.* Delaunay). Each simplex σ of S is embedded in $\mathbb{X}$ as the convex hull of its vertices W_σ. The probability density function of the GSC is the convex combination (with prior π_σ) of the convolution of the noise density g over each embedded simplex W_σ: $\forall x \in \mathbb{X},\ p(x, S, W, \theta) = \sum_{\sigma \in S} \frac{\pi_\sigma}{|W_\sigma|} \int_{W_\sigma} g(x, w, \theta) dw$, with $\sum_{\sigma \in S} \pi_\sigma = 1$ and typically $g(x, \mu, \theta)$ is a density function with mean μ and variance θ. The GSC reduces to the standard GMM when $S = W$ (0-skeleton), and g is the Gaussian density function.

The parameters w, θ and π_σ adjust the density model p to the data sample, and are typically optimized using the Expectation-Maximization technique [11]. The π_σ parameters are used to define a filtration of S [4, 12, 5] to compute a persistence diagram, or thresholded to carve out a sub-complex S^* of S whose topology accounts for that of M. The number K of landmarks impacts the size of S hence the "possible richness" of the topological model S^* of M, and so controls the model over-fitting. However, there is a missing theoretical link between the topology of S^* and standard statistical indicators used to select "good" density models p, *e.g.* the Bayesian Information Criterion [13] used to find an optimal K, that could inform on the topological validity of S^* with respect to the target topology of M.

3 Research direction for Validation

Our overall objective is to introduce a paradigm shift in topological data analysis, where more focus is put on selecting meaningful simplicial complexes that generalize well to unseen data.

Mapper Unsupervised methods can be validated through a direct measure of how good a model fits the data or by comparing the results from a training dataset to the results from a validation dataset. Direct measures for mapper could be defined as a combination of cluster quality and consistency of clusters, but we found no previous work in this direction. On the other hand, Reeb graphs have been compared using many different metrics, for example the Gromov-Hausdorff distance [14], functional distortion distance [15], interleaving distance [16], edit distance [17] and intrinsic distances [18]. In addition, mapper complexes are essentially covers, and methods to compare clusters may be extended to also compare covers as well. For example one may consider an extension of an information-theoretic approach to cluster distances [19]. Some of these distances have been applied to compare mapper graphs. However, it is not clear how these distances depend on the data sizes and on the chosen cover of $\mathbb{R}$. This raises a question about how we can reliably validate mapper graphs when the training and test data are of different sizes and how we select optimal hyperparameters for mapper in that setting. Good validation measures should be independent of data sizes, hyperparameters and data noise distributions.

Generative Simplicial Complex In the GSC model, beyond intuition and empirical evidence in simple cases [4, 12, 5], there is no theoretical proof that the statistical inference process which generates S^* from a "good" density model p of the data, is also relevant for selecting a "good" topological model of the latent manifolds M. In order to analyze the topological validity of the GSC model, we plan to study all the factors at play beyond the data density function, like the intrinsic dimension and curvature of the latent manifolds, or the quantity of data available to estimate the parameters of the GSC, but also the likelihood objective function, the selection criteria and inference techniques used for this estimation.

References

[1] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, Nov 2002.

[2] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[3] Harish Doraiswamy, Julien Tierny, Paulo J. S. Silva, Luis Gustavo Nonato, and Claudio Silva. Topomap: A 0-dimensional homology preserving projection of high-dimensional data. In *Proc. of IEEE VIS conference. IEEE Transactions on Visualization and Computer Graphics*, 2020.

[4] Michaël Aupetit. Learning topology with the generative gaussian graph and the EM algorithm. In *Advances in Neural Information Processing Systems 18*, pages 83–90, 2005.

[5] Maxime Maillot, Michaël Aupetit, and Gérard Govaert. A generative model that learns betti numbers from a data set. In *20th European Symposium on Artificial Neural Networks, ESANN 2012, Bruges, Belgium, April 25-27, 2012*, 2012.

[6] Pek Y. Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkhanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson, and Gunnar Carlsson. Extracting insights from the shape of complex data using topology. *Sci Rep*, 3:1236, 2013.

[7] Julio-Omar Palacio-Niño and Fernando Berzal. Evaluation metrics for unsupervised learning algorithms, 2019.

[8] M. Nicolau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. U.S.A.*, 108(17):7265–7270, Apr 2011.

[9] A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and R. Rabadan. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.*, 35(6):551–560, 06 2017.

[10] Mathieu Carrière, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for mapper. *Journal of Machine Learning Research*, 19(12):1–39, 2018.

[11] Halima Bensmail, Gilles Celeux, Adrian E. Raftery, and Christian P. Robert. Inference in model-based cluster analysis. *Stat. Comput.*, 7(1):1–10, 1997.

[12] Pierre Gaillard, Michaël Aupetit, and Gérard Govaert. Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing*, 71(7-9):1283–1299, 2008.

[13] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6:461–464, 1978.

[14] David A. Edwards. The structure of superspace. In *Studies in topology (Proc. Conf., Univ. North Carolina, Charlotte, N. C., 1974; dedicated to Math. Sect. Polish Acad. Sci.)*, pages 121–133. Academic Press, New York, 1975.

[15] Ulrich Bauer, Xiaoyin Ge, and Yusu Wang. Measuring distance between Reeb graphs [extended abstract]. In *Computational geometry (SoCG'14)*, pages 464–473. ACM, New York, 2014.

[16] Vin de Silva, Elizabeth Munch, and Amit Patel. Categorified Reeb graphs. *Discrete Comput. Geom.*, 55(4):854–906, 2016.

[17] Barbara Di Fabio and Claudia Landi. The edit distance for Reeb graphs of surfaces. *Discrete Comput. Geom.*, 55(2):423–461, 2016.

[18] Mathieu Carrière and Steve Oudot. Local equivalence and intrinsic metrics between reeb graphs. *CoRR*, abs/1703.02901, 2017.

[19] Marina Meilă. Comparing clusterings - an information based distance. *J. Multivariate Anal.*, 98(5):873–895, 2007.