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
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 \( M \subset \mathbb{R}^d \) 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 \( \delta \). 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 hyperparameters 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 \( \sigma \in \mathbb{R}^K \) (e.g. the component means of a GMM) and a simplicial complex \( S \) of \( W \) (e.g. Delaunay). Each simplex \( \sigma \in \tau \) is embedded in \( \mathbb{R} \) as the convex hull of its vertices \( W_{\sigma} \). The probability density function of the GSC is the convolution of the noise density \( g \) over each embedded simplex \( W_{\sigma} \). The density is put on selecting meaningful simplicial complexes that generalize well to unseen data.

\[ p(x, S, W, \theta) = \sum_{\sigma \in \tau} p(\pi_{\sigma}, \chi) \int_{W_{\sigma}} g(x, w, \theta) dw, \]

where \( \pi_{\sigma} \) is a density function with mean \( \mu \) and variance \( \theta \). The GSC reduces to the standard GMM when \( S = W \) (0-skeleton), and \( g \) is the Gaussian density function.

The parameters \( w, \theta \) and \( \pi_{\sigma} \) adjust the density model \( p \) to the data sample, and are typically optimized using the Expectation-Maximization technique \([11]\). The \( \pi_{\sigma} \) 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


