A Probabilistic Perspective on Persistence Homologies

SAMSI 2007 working group

October 12, 2007

Introduction

Topological persistence (Edelsbrunner *et al.*, 2002) provides an efficient technique for discovering the underlying topological structure of data. The objective for extending topological ideas such as persistence homologies to a probabilistic setting is driven by three strongly related reasons: quantification of uncertainty, estimates of the number of samples required to achieve a particular resolution of the estimate, and removal of outliers and denoising.

The basic probabilistic perspective is that a point cloud is a sample of n points drawn from a distribution on a compact subset of Euclidean space, in our formulation a smooth manifold or a smooth manifold with noise. This sample is a random object and so is the persistence diagram (Cohen-Steiner *et al.*, 2007) constructed by a filtration based on the point cloud. The corresponding underlying object is the manifold itself which also has a persistence diagram. The basic questions from a probabilistic viewpoint are
(1) Scaling – In terms of the number of samples and filtration what is the deviation of the persistence diagrams from a sample and the manifold. Consistency – does this deviation goto zero with an infinite sample size. Is there central limit theorem or large deviation formulation of this problem – the natural scale of the distribution of errors between the diagrams.
(2) Credible (confidence) intervals – Given the above notion of scale can one provide a confidence interval for each point of the persistence diagram constructed from a point cloud. Can one say that each point in this diagram is with high probability ϵ-near a point in the persistence diagram of the manifold. The credible interval computation and sample size computation are fundamentally linked since ϵ is a function of n via large deviation results from the scale analysis.
(3) Denoising – A standard problem is statistics is to model

$$\text{observations} = \text{object} + \text{noise},$$

where in this case the object is the manifold and our summary statistic from the observation is the persistence diagram. The question is can we filter noisy samples based on operations on the persistence diagram ?

We develop two aspects of probabilistic modeling for persistence homologies: a theoretical worst case result and algorithmic/statistical procedures to provide estimates of variation or uncertainty.

Theoretical results

The setting we consider is a $p_{\mathcal{M}}$-dimensional compact manifold $\mathcal{M}$ embedded in a p-dimensional Euclidean space. We obtain a sample of n points $X_n = \{x_1, ..., x_n\}$ drawn independently from a uniform distribution on $\mathcal{M}$.

For the manifold a distance function can be used to construct the persistence diagram. Persistence diagrams of the point cloud X_n can be constructed from the distance function on the points in the point cloud using simplicial complexes such as the α complex (Edelsbrunner *et al.*, 2002; Zomorodian and Carlsson, 2004). Our objective will be to make probabilistic statements as to how close the persistence diagram of X_n is to the persistence diagram of $\mathcal{M}$. This is very much in the spirit of the probabilistic approximately correct (PAC) framework (Valiant, 1984).

Theorem 1 (Probabilistic Persistent Homology Inference) *Define $D(\mathcal{M})$ and $D(X_n)$ as the persistence diagram of the manifold and a sample from the manifold. Given X_n, with probability at least $1-\delta$*

the points in $D(X_n)$ which are at least ϵ away from the diagonal represent actual persistent homological features. Moreover, for every one of those points $p \in D(X_n)$, the point of $D(\mathcal{M})$ representing the true persistent homology class lies in an ϵ-box around p. Under mild assumptions

$$\epsilon = \mathcal{O}\left(\log(\frac{1}{\delta})n^{-1/p_{\mathcal{M}}}\right).$$

Corollary 1 (Probabilistic Homology Inference Theorem) *Given X_n if for the same ϵ as above*

$$\epsilon < \frac{hfs(\mathcal{M})}{3}$$

then with probability at least $1-\delta$ for all dimensions i the dimension of $H_i(\mathcal{M})$ is the same as the number of points in $D_i(X_n)$ which lie above and to the left of the point $(\epsilon, 2\epsilon)$ where $hfs(\mathcal{M})$ is the homological feature size of the manifold, and H_i and D_i are the homology group and persistence diagram for dimension i.

The second result improves upon a homology inference result by Niyogi *et al.* (2007) by replacing the reach of the manifold with the homological feature size (hfs). The reach is a strict upper bound on the hfs that can be arbitrarily loose. From this perspective our theorem is much stronger. However, the results in Niyogi *et al.* (2007) apply to homotopy while our results do not. This is not a serious shortcoming since homotopy is typically not computable from data.

Algorithms to estimate uncertainty

The above theory serves as an initial attempt to bound uncertainty of persistence diagrams. The bounds in these theorems like most PAC bounds are useless in practice: they are too loose.

We are exploring two sample based ideas on simulated data to provide practical intervals of uncertainty. The two ideas are bootstrap methods to resample the data and provide resampled persistence diagrams and Bayesian approaches that provide persistence diagrams based on models of the data and posterior probabilities for these models. We have preliminary implementation of the bootstrap method and can average of persistence diagrams in a principled way.

Open problems
1) Tightness of our theoretical bounds.
2) Generalization of these bounds to compact subsets in Euclidean space (Chazal *et al.*, 2007).
3) A bootstrap central limit theorem for persistence homologies.
4) Mathematical understanding of posterior distribution on persistence diagrams or algebraic models of topological structure.

References

Chazal, F., Cohen-Steiner, D. and Lieutier, A. (2007). A sampling theory for compacts in euclidean space,.

Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120.

Edelsbrunner, H., Letscher, D. and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Compt Geom* **28** 511–533.

Niyogi, P., Smale, S. and Weinberger, S. (2007). Finding the homology of submanifolds with high confidence from random samples. *To appear in Discrete Compt. Geom.* .

Valiant, L. G. (1984). A theory of the learnable. In *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of computing*, 436–445. ACM Press, New York, NY, USA. doi: http://doi.acm.org/10.1145/800057.808710.

Zomorodian, A. and Carlsson, G. (2004). Computing persistent homology. In *SCG '04: Proceedings of the twentieth annual symposium on Computational geometry*, 347–356. ACM Press.