COMPUTING PERSISTENT HOMOLOGY UNDER RANDOM PROJECTION

Karthikeyan Natesan Ramamurthy,[1] *Kush R. Varshney,*[1] *and Jayaraman J. Thiagarajan*[2]

[1]IBM Thomas J. Watson Research Center
[2]Lawrence Livermore National Laboratory

ABSTRACT

Random projection is a tried-and-true technique in signal processing for reducing sensing complexity while maintaining acceptable performance of downstream processing tasks. In this paper, we investigate random linear projection of point clouds followed by topological data analysis for computing persistence diagrams and Betti numbers. In this first empirical study of its kind in the literature, we find that Betti numbers can be recovered accurately with high probability after random projection up to a certain reduced dimension but then the probability of recovery decreases to zero. We further investigate how the mean of the persistence diagrams from several random projections can be used favorably in Betti number recovery. Our empirical study includes both synthetic data as well as real-world range image and respiratory audio data.

Index Terms— Betti numbers, persistence diagrams, random projection, topological signal processing

1. INTRODUCTION

An important aspect of the data revolution that is happening today is that huge quantities of various types of data are being produced at an ever accelerating rate. This necessitates the development of novel tools to analyze and understand this data. While geometric methods focus on measuring and visualizing the *size* of data, topological approaches help us qualitatively understand the *shape* of data and provide high-level summaries. For example, in Fig. 1(a), the point cloud sampled from the circle can be topologically quantified as having one connected component, and one 1-dimensional 'hole.' Topological data analysis is now starting to be used in a variety of challenging signal processing applications ranging from sensor networks, to respiratory disease diagnosis, to social network analysis.

The data sampled from shapes in a high-dimensional space equipped with a distance function are referred to as *point clouds*; several approaches have been developed to infer the topology of the shape from the point cloud. However, sensing and analyzing large, high-dimensional, noisy point clouds possibly ridden with outliers poses unique challenges. In this paper, we are interested in computing the *topological features* of data from reduced-dimensional measurements obtained with random projections [1]. This becomes crucial largely from the sensing perspective, and also the storage perspective in some scenarios, when the data is high-dimensional.

Specifically, we perform topological inference on point clouds using random linear measurements obtained from them. Random projections have been shown to preserve isometry in the data approximately [1]. Recently, this has been used to prove that they also approximately preserve persistent homology [2]. However, so far there has been no empirical analysis reported in the literature on the effect of random projections on obtaining persistence. We perform an extensive empirical analysis, and furthermore, we compute means of persistence diagrams obtained from multiple random projections [3]. We show experimentally that by using mean persistence from multiple projected versions of the data, we are able to recover homology cycles better than when using persistence from a single projection.

Part of this work was performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-650593.

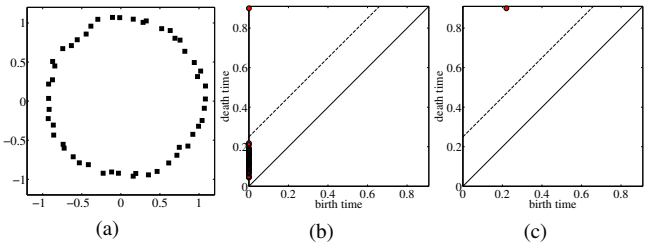

Fig. 1. (a) Noisy point cloud from a circle in 2-dimensional space, (b) persistence diagram for H_0, (c) persistence diagram for H_1.

2. PERSISTENT HOMOLOGY AND BETTI NUMBERS

Let us denote the point cloud with T samples as $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$. A simplicial complex $\mathcal{S}$ can be built to approximate the topological structure of $\mathcal{X}$. We consider the samples $\mathcal{X}$ as vertices, and progressively add edges, triangles and high-dimensional cells creating higher dimensional simplices that are connected to each other. The simplicial complex is formally defined as the pair (V, Σ), where V is the set of vertices, and Σ is a family of non-empty subsets of V, where each subset denotes a simplex, with the condition that $\sigma \in \Sigma$ and $\tau \subseteq \sigma$, implies that $\tau \in \Sigma$. Using the simplicial complex $\mathcal{S}$, we can compute the *Betti numbers*, $\beta_k(\mathcal{S})$, which are the number of k-dimensional holes of the complex. For example, if $\mathcal{S}$ is obtained from a circle as in Fig. 1, we have $\beta_0 = 1$, which is the number of connected components (or 0-dimensional holes), and $\beta_1 = 1$, which is the number of 1-dimensional holes. Clearly, $\beta_k = 0$ for $k > 1$. Formally. the Betti number $\beta_k(\mathcal{S})$ is the dimension of the k^{th} homology group of the complex, $H_k(\mathcal{S})$ [4].

There are various approaches to compute simplicial complexes from $\mathcal{X}$. In the Čech complex with the scale parameter ϵ, denoted as a Čech$(\mathcal{X}, \epsilon)$, a simplex is created between a set of vertices G if and only if there is a non-empty intersection of the closed Euclidean balls $B(x_i, \epsilon/2), \forall i \in G$. In the Vietoris-Rips (VR) complex, $VR(\mathcal{X}, \epsilon)$, a simplex is created if and only if the Euclidean distance between vertices belonging to every edge is less than ϵ. Hence a VR complex is completely determined by the distance between the points in $\mathcal{X}$.

Since both these complexes consider all the T points, they result in a large number of simplices, thereby increasing the computational complexity of topological inference. The witness complex proposed in [5], overcomes this problem by computing the complex with only L landmark points, where $L \ll T$. In this complex, an edge is created between two landmarks if and only if there exists a witness point in $\mathcal{X}$ whose two closest neighbors are those landmarks. In the *lazy* witness complex, higher dimensional simplices are created based on the 1-skeleton. Variants of the witness complex that depend on a scale parameter ϵ, also exist [5]. Note that the scale parameter, ϵ, will also be referred to as *time*.

The homological inference depends on the scale (time) at which the complexes are constructed. Therefore, it is important to identify the stable topological features across scales, i.e. that are *persistent*. The persistent features provide a summary of the homological information for many different values of ϵ at once. Considering only the holes or the homology cycles, we can obtain *persistence diagrams* for each Betti number that denote the birth and death times of each homology cycle. Referring again to the example in Fig. 1, it can be seen that the homology cycles that persist for a long time represent the stable topological features. The diagram for Betti k is denoted as P_k. Obtaining simplicial complexes at various scales and computing persistence diagrams involves a lot of computations even for a moderate number of points. Therefore, the complexity is limited by fixing the maximum scale as t_{max} and choosing a small set of samples, either randomly or adaptively from the point cloud to obtain persistence diagrams. When the shapes are submanifolds, [6] discusses the conditions under which the homology can be inferred with high confidence. Methods for obtaining linear size filtered simplicial complexes, such that its persistence diagram is a good approximation to that of VR filtration also exist [7].

3. STATISTICS ON PERSISTENCE DIAGRAMS

Since the process of constructing simplicial complexes and computing persistence diagrams involves sampling and numerical approximation, there is a need to eliminate the noise introduced in this process to improve topological inference. For example, in Fig. 1(b), clearly the homology cycles showing up close to the diagonal, below the dotted noise threshold line, are born and die in a short time, and hence are not significant. Eliminating the topological noise in this case is equivalent to preserving only the homology cycles whose difference between death and birth times are greater than a threshold. Such topological signal processing strategies become important when we attempt to extract and use topological features from real-world data. Understanding the characteristics of the space of persistence diagrams will help us develop more sophisticated topological signal processing and statistical analysis methods.

In [8], the authors prove a stability property for the persistence diagrams. In practical terms, we can expect that persistence diagrams obtained from two point clouds of two shapes that are close to each other will be close. Furthermore, the space of persistence diagrams itself is a metric space endowed with the Wasserstein metric, and it also allows for the definition of probability measures which can be used to compute various statistical quantities [9]. In particular, the authors in [9] show that the Fréchet mean for a finite set of persistence diagrams always exists. In [10], this idea is extended to define a probabilistic Fréchet mean, that varies continuously for continuously varying diagrams. Confidence intervals for rejecting the noise from the signal in persistence diagrams are introduced in [11, 12].

4. PROPOSED APPROACH

In order to examine the effect of random projections, we obtain persistence diagrams from a randomly projected version, $R(\mathcal{X})$, of the point cloud $\mathcal{X}$. Each sample $R(x_i) \in R(\mathcal{X})$ is the random projection of $x_i \in \mathcal{X}, \forall i \in \{1, \ldots, T\}$, where R is the dimensionality reducing random projection that preserves approximate isometry in the Euclidean space. When the sensing budget is limited, random projections can be beneficial in capturing essential topological information at a low sensing and storage cost. Complexes can be directly computed on $R(\mathcal{X})$, and since these depend only on the Euclidean distance between the samples, the persistent homology will be approximately preserved. However, we note that the computational complexity reduction achieved with random projections is not significant, since computing the distance matrix is not the dominant complexity step while obtaining the Betti numbers.

We will perform an empirical analysis of the probability of accurate recovery of Betti numbers from the random projected point cloud. Furthermore, we will also compute the Fréchet mean of multiple persistence diagrams, $P_k^{(m)}, m = \{1, \ldots, M\}$, obtained from the random projected point clouds, $R_m(\mathcal{X}), m = \{1, \ldots, M\}$, for the homology group k, and analyze the effectiveness of recovering β_k from the mean diagram $\hat{P}_k$. The mean persistence is obtained as

$$\hat{P}_k = \underset{P_k}{\operatorname{argmin}} \sum_{m=1}^{M} W_2^2(P_k, P_k^{(m)}) \tag{1}$$

where $W_2(.,.)$ is the 2-Wasserstein distance between the two diagrams, and the minimum value of the objective is the variance. Computing the 2-Wasserstein distance involves pairing each point in one of the diagrams with a point in the other, and summing the squared Euclidean distances between the paired points. One of the paired points can be on the diagonal, since it contains homology cycles that are born and die at the same time.

The outline of the algorithm to compute the mean persistence is as follows [3]:

1. Initialize $\hat{P}_k$ to one of the M diagrams, randomly.
2. Compute the optimal pairing between the points in $\hat{P}_k$ to each of the diagrams, $P_k^{(m)}, m = \{1, \ldots, M\}$, using the Hungarian algorithm.
3. Update each point in $\hat{P}_k$, with the arithmetic mean over its pairs in all the diagrams. Note that each point in $\hat{P}_k$ is paired with a unique point on each of the diagrams, $\{P_k^{(m)}\}_{m=1}^{M}$, and one of the points in the pair can be on the diagonal.
4. Continue steps 2 and 3 until the objective in (1) converges.

5. EXPERIMENTS AND RESULTS

5.1. Data Sets and Experimental Setting

We conduct the experiment using three data sets: one synthetic data set and two real-world data sets. The synthetic data set consists of a point cloud sampled from the unit torus in four dimensions with additive uniform noise drawn from the range $[0, 1/4]$. To this $4-$dimensional data, we append 46 dimensions of uniform noise drawn from the range $[-1/20, 1/20]$. This data set contains 10,000 points. The true Betti numbers for this data set are: $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$, and all others zero. The first real-world data set has 15,000 vectors of 25 dimensions obtained from range image patches, i.e. image patches indicating distance from the observer to objects in

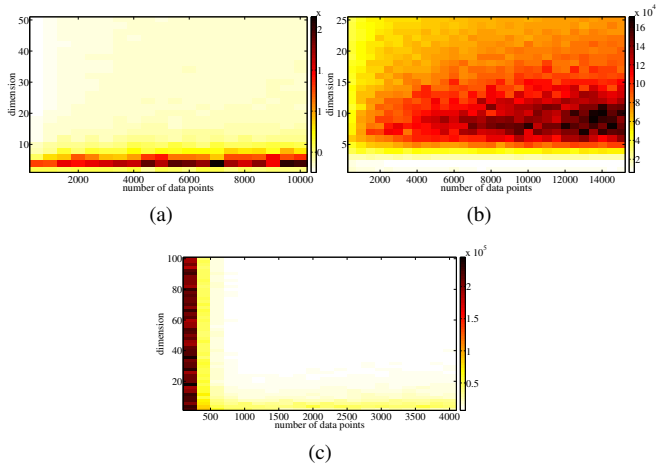

Fig. 2. Number of simplices in (a) torus, (b) range image, and (c) wheezing audio data sets averaged over 50 trials.

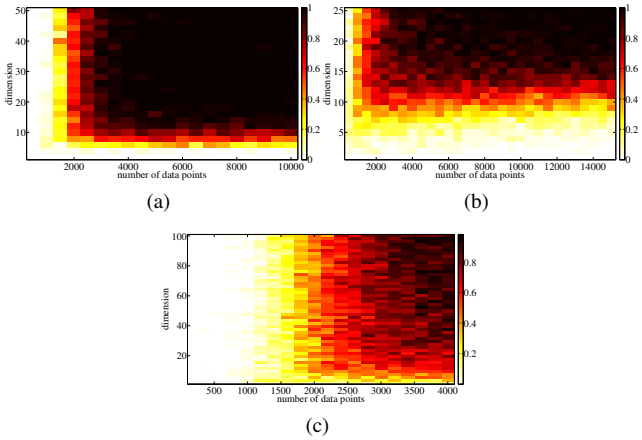

Fig. 3. Probability of correct identification of Betti numbers in (a) torus, (b) range image, and (c) wheezing audio data sets estimated using 50 trials.

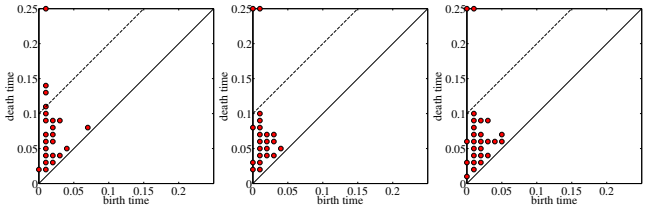

Fig. 4. Persistence diagrams of H_1 of random trials of the torus data set at reduced dimension 30 and 2000 data points.

the scene, as described in [13]. The Betti numbers for the range image patches are found to be $\beta_0 = 1, \beta_1 = 1$, and all others zero. The second real-world data set is a delay embedding from the audio domain. It is derived from a clip of a wheezing patient. This data set is 100-dimensional with 4,000 data points and ideally should have Betti numbers $\beta_0 = 1, \beta_1 = 1$, and all others zero [14].

In the experiments, we sweep over two parameters: the reduced dimension via random linear projection, and the number of data points. The random projection matrices are drawn uniformly from the Stiefel manifold [15], and the data points are randomly sampled without replacement from the full data set. We use lazy witness complexes for computing persistence diagrams with 75 landmarks chosen in a sequential min-max way starting from a random initialization. We compute persistence diagrams for 50 random trials at each pair of reduced dimension and number of data points. For the first two data sets, we set $t_{\max} = 0.25$ and for the third, $t_{\max} = 0.6$.

5.2. Results and Discussion

The average number of simplices that are created across the trials are plotted for the three data sets in Fig. 2. The number of simplices has a direct relationship with computation time. In the torus and range image data sets, we see that the maximum number of simplices occurs at a high number of points and intermediate reduced dimension. In the wheezing audio data set, many simplices are found at very low number of points and the pattern of the other two data sets occurs in the remainder of the space. Computational complexity is an important but secondary consideration, since random projection affects only the computation of distance matrices, which is not the dominant complexity step in obtaining persistence diagrams.

Fig. 3 shows the fraction of the trials in which we recover the true Betti numbers, which is a sample estimate of the probability of recovery. We calculate the Betti numbers by first capping the death time of any homology cycle at $t_{\max}$, and then counting how many death time–birth time differences are greater than the noise threshold: 0.1 for the first two data sets and 0.4 for the third data set. A larger threshold is used for the wheezing data set because it is noisier. The results in Fig. 3 remain qualitatively similar for large ranges of noise thresholds. As would be expected, the further to the top right corner of the plots we go, which involves ever greater sensing resources, the probability of correct recovery goes to one. With the synthetic torus data set, the probability of recovery exhibits a much more threshold-like behavior in both parameters, whereas in the real-world data sets, the degradation in this probability is much more gradual. Clearly, even without sacrificing any probability of correct Betti number recovery, it is possible to reduce a little bit of sensing complexity. Moreover, if one is willing to sacrifice some of the probability of correct Betti number recovery (from a single trial), then one can reduce the sensing complexity even further. It turns out that the highest computational complexity in determining persistence diagrams, as revealed by the number of simplices, tends to be at the points where the recovery probability is around 0.3 to 0.5, which is probably too low for most applications anyway.

A common modus operandi in signal processing is to take sample averages for noise reduction. Here we examine the first homology group of the torus data set which should have $\beta_1 = 2$. Fig. 4 shows the persistence diagrams from the first three of fifty trials used in estimating the values in Fig. 2(a) and Fig. 3(a) for a set of parameters that yields approximately 50% recovery of the true set of Betti numbers. The first sample only has one point at death time $t_{\max}$ whereas the other two samples have two each. The first sample has two other points above the noise threshold whereas the others have none. Thus, the estimate of β_1 is three from the first sample and is two from the others. The mean persistence diagram from the first two samples, shown in Fig. 5, recovers the two salient points at $t_{\max}$ but continues to have one other point above the threshold. In mean persistence diagrams from larger numbers of trials, this extra point remains when averaging three trials, is right at the threshold for four samples, comes up above the threshold for eight trials, and

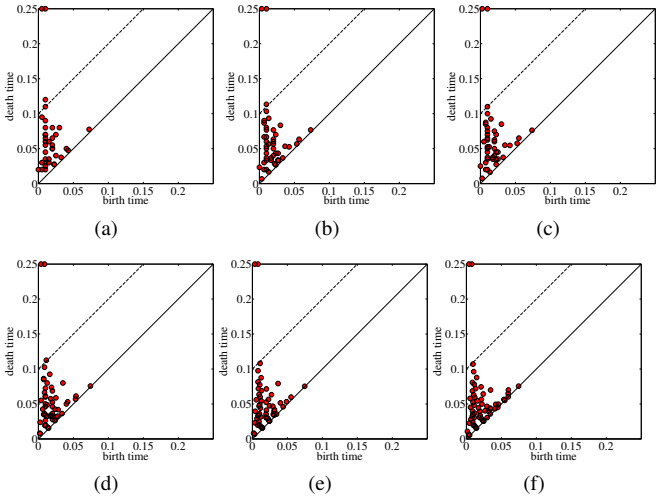

Fig. 5. Mean persistence diagrams of H_1 of first (a) 2, (b) 3, (c) 4, (d) 8, (e) 20, and (f) 50 random trials of the torus data set at reduced dimension 30 and 2000 data points.

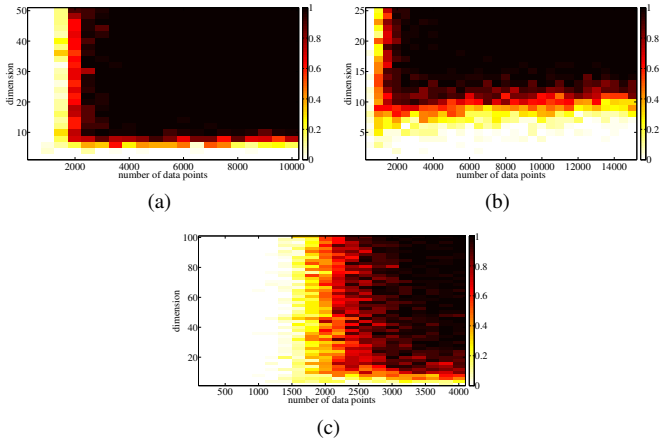

Fig. 6. Probability of correct identification of Betti numbers from mean persistence diagram of 3 trials in (a) torus, (b) range image, and (c) wheezing audio data sets, estimated using 50 random samples of 3 persistence diagrams each.

is correctly below the threshold for twenty and fifty trials. This example illustrates that computing mean persistence diagrams is able to improve Betti number recovery by eventually lessening the impact of errors such as in the first sample. To illustrate this further, we show plots of the probability for accurately recovering Betti numbers from mean persistence diagrams of three trials each in Fig. 6. With all three data sets, the range of parameters over which we are able to recover the Betti numbers with high probability is enlarged via mean persistence diagrams. The average probability of recovery across the parameter settings is 0.738, 0.610, and 0.455 for the three datasets with mean persistence (Fig. 6), whereas it is 0.718, 0.566, and 0.359 with individual persistences (Fig. 3).

6. CONCLUSION

Using empirical experiments, we have shown that random projections can be used for efficient sensing and storage, when the ultimate objective to obtain persistence diagrams that describe the topology of the data. To the best of our knowledge, this is the first work that uses random projections on real datasets for topological signal processing, and demonstrates the benefits in obtaining mean persistence diagrams for removing topological noise. Future research directions include incorporating prior knowledge of the geometry or the rough topological structure to further reduce the sensing complexity and using robust statistics on persistence diagrams to eliminate the spurious homological cycles that may show up in persistence diagrams.

7. REFERENCES

[1] S. S. Vempala, *The Random Projection Method*. Providence, RI: American Mathematical Society, 2004.

[2] D. R. Sheehy, "The persistent homology of distance functions under random projection," in *Proc. Symp. Comput. Geom.*, Kyoto, Japan, Jun. 2014.

[3] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer, "Fréchet means for distributions of persistence diagrams," Available at http://arxiv.org/pdf/1206.2790, Mar. 2013.

[4] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc.*, vol. 46, no. 2, pp. 255–308, Apr. 2009.

[5] V. de Silva and G. Carlsson, "Topological estimation using witness complexes," in *Proc. Eurographics Symp. Point-Based Graphics*, Zurich, Switzerland, Jun. 2004, pp. 157–166.

[6] P. Niyogi, S. Smale, and S. Weinberger, "Finding the homology of submanifolds with high confidence from random samples," *Discrete Comput. Geom.*, vol. 39, no. 1–3, pp. 419–441, 2008.

[7] D. R. Sheehy, "Linear-size approximations to the Vietoris–Rips filtration," *Discrete Comput. Geom.*, vol. 49, no. 4, pp. 778–796, Jun. 2013.

[8] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, "Stability of persistence diagrams," *Discrete Comput. Geom.*, vol. 37, no. 1, pp. 103–120, 2007.

[9] Y. Mileyko, S. Mukherjee, and J. Harer, "Probability measures on the space of persistence diagrams," *Inverse Probl.*, vol. 27, no. 12, p. 124007, Dec. 2011.

[10] E. Munch, P. Bendich, K. Turner, S. Mukherjee, J. Mattingly, and J. Harer, "Probabilistic Fréchet means and statistics on vineyards," Available at http://arxiv.org/pdf/1307.6530, 2013.

[11] P. Bubenik, "Statistical topological data analysis using persistence landscapes," Available at http://arxiv.org/pdf/1207.6437, Dec. 2013.

[12] S. Balakrishnan, B. Fasy, F. Lecci, A. Rinaldo, A. Singh, and L. Wasserman, "Statistical inference for persistent homology," Available at http://arxiv.org/pdf/1303.7117, Mar. 2013.

[13] H. Adams and G. Carlsson, "On the nonlinear statistics of range image patches," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 110–117, 2009.

[14] S. Emrani, T. Gentimis, and H. Krim, "Persistent homology of delay embeddings," Available at http://arxiv.org/pdf/1305.3879, Jul. 2013.

[15] Y. Chikuse, *Statistics on Special Manifolds*. New York, NY: Springer-Verlag, 2003.