Dimension Detection with Local Homology

Tamal K. Dey* Fengtao Fan† Yusu Wang‡

Abstract

Detecting the dimension of a hidden manifold from a point sample has become an important problem in the current data-driven era. Indeed, estimating the shape dimension is often the first step in studying the processes or phenomena associated to the data. Among the many dimension detection algorithms proposed in various fields, a few can provide theoretical guarantee on the correctness of the estimated dimension. However, the correctness usually requires certain regularity of the input: the input points are either uniformly randomly sampled in a statistical setting, or they form the so-called (ε, δ)-sample which can be neither too dense nor too sparse.

Here, we propose a purely topological technique to detect dimensions. Our algorithm is provably correct and works under a more relaxed sampling condition: we do not require uniformity, and we also allow Hausdorff noise. Our approach detects dimension by determining local homology. The computation of this topological structure is much less sensitive to the local distribution of points, which leads to the relaxation of the sampling conditions. Furthermore, by leveraging various developments in computational topology, we show that this local homology at a point z can be computed *exactly* for manifolds using Vietoris-Rips complexes whose vertices are confined within a local neighborhood of z. We implement our algorithm and demonstrate the accuracy and robustness of our method using both synthetic and real data sets. Missing details from this abstract can be found in the full version of this paper [7].

1 Introduction

Learning about a manifold embedded in $\mathbb{R}^d$ from its point data is a key problem in various manifold learning applications. Most times, the intrinsic dimension of the manifold M is one of the simplest, yet still very important, quantities that one would like to infer from input data. Therefore, there has been considerable research in dimension estimation for manifolds. Under the statistical setting, different approaches estimate the manifold dimension based on the growth rate of the volume (or some analog of it) of an intrinsic ball [9, 11]. They generally assume that the input points are sampled from some probabilistic distribution supported on the hidden manifold. In the computational geometry community, there are provable dimension detection algorithms which all require (ε, δ)-sampling condition that points are both ε-dense and δ-sparse. Cheng et al. [5] developed an improved algorithm from them to tolerate a small amount of Hausdorff noise (of the order ε^2 times the local feature size). More recently, Cheng et al. [4] proposed an algorithm to estimate the dimension by detecting the so-called *slivers*. This algorithm assumes that the input points are sampled from the hidden manifold using a Poisson process without noise.

In this paper we develop a dimension detection method based on the topological concept of local homology. The idea of using local homology to study stratified spaces from sampled points was first proposed by Bendich et al. [1] and further explored in [2]. Both of them used Delaunay triangulations in their algorithms. In a recent paper [15], Skraba and Wang proposed to approximate the multi-scale representations of local homology using families of Rips complexes. In the same spirit, we show that the dimension of a manifold can also be deduced from Rips complexes using the local homology.

Our results. Given a smooth m-dimensional manifold M embedded in $\mathbb{R}^d$, the local homology group $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ at a point $z \in \mathsf{M}$ is isomorphic to the reduced homology group of a m-dimensional sphere, that is $\mathsf{H}(\mathsf{M}, \mathsf{M} - z) \cong \tilde{\mathsf{H}}(\mathbf{S}^m)$. Hence, given a set of noisy sample points P of M, we aim to detect the dimension of M by estimating $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ from P. Specifically, we assume that P is an ε-sample[1] of M in the sense that the Hausdorff distance between P and M is at most ε. Our main result is that by inspecting two nested neighborhoods around a sample point $p \in P$ and considering certain relative homology groups computed from the Rips complexes induced by points within these neighborhoods, one can recover the local homology *exactly*; see Theorem 13. This in turn provides a provably correct dimension-detection algorithm for an ε-sample

*Dept. of Computer Science & Engineering, Ohio State University, tamaldey@cse.ohio-state.edu

†Dept. of Computer Science & Engineering, Ohio State University, fanf@cse.ohio-state.edu

‡Dept. of Computer Science & Engineering, Ohio State University, yusu@cse.ohio-state.edu

[1] Note that this definition of ε-sample allows points in P to be ε distance off the manifold M. Our ε-sampling condition is with respect to the *reach* of M while that used in [4, 5, 8, 10] is with respect to local feature size and thus adaptive.

P of a hidden manifold M when ε is small enough.

Compared with previous provable results in [4, 5, 8, 9, 10, 13], our theoretical guarantee on the estimated dimension is obtained with a more relaxed sampling condition on P. Specifically, there is no uniformity requirement for the sample points P, which was required by all previous dimension-estimation algorithms with theoretical guarantees: either in the form of a uniform random sampling in the statistical setting [4, 9, 13] or the (ε, δ)-sampling in the deterministic setting [5, 8, 10]. We also allow larger amount of noise (ε vs. ε^2 as in [4]). Such a relaxation in the sampling condition is primarily made possible by considering the topological information, which is less sensitive to the distribution of points compared to the approaches based on local fitting.

In Section 6, we provide preliminary experimental results of our algorithm on both synthetic and real data. For synthetic data our method detects the right dimension robustly. For real data some of which are laden with high noise and undersampling, not all points return the correct dimension. But, taking advantage of the fact that local homology is trivial in all but zero and intrinsic dimension of the manifold, we can eliminate most false positives and estimate the correct dimension from appropriately chosen points.

Finally, we remark that similar to the recent work in [15], our computation of local homology uses the Rips complex, which is much easier to construct than the ambient Delaunay triangulation as was originally required in [1]. Different from [15], we aim to compute $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ *exactly* for the special case when M is a manifold, while the work in [15] *approximates* the multi-scale representations of local homology (the persistence diagram of certain filtration) for more general compact sets. The goals from these two works are somewhat complementary and the two approaches address different technical issues.

Remark. The missing details and proofs from this abstract can be found in the full version of the paper [7].

2 Preliminaries and Notations

Manifold and sample. Let M be a compact smooth m-dimensional manifold without boundary embedded in an Euclidean space $\mathbb{R}^d$. The *reach* $\rho(\mathsf{M})$ is the minimum distance of any point in M to its medial axis. A finite point set $P \subset \mathbb{R}^d$ is an *ε-sample* of M if every point $z \in \mathsf{M}$ satisfies $d(z, P) \leq \varepsilon$ and every point $p \in P$ satisfies $d(p, \mathsf{M}) \leq \varepsilon$; in other words, the *Hausdorff distance* between P and M is at most ε.

Balls. An Euclidean closed ball with radius r and center z is denoted $B_r(z)$. The open ball with the same center and radius is denoted $\mathring{B}_r(z)$ and its complement $\mathbb{R}^d \setminus \mathring{B}_r(z)$ is denoted $B^r(z)$.

Homology. We denote the i-th dimensional homology group of a topological space X as $\mathsf{H}_i(X)$. We drop i and write $\mathsf{H}(X)$ when a statement holds for all dimensions. We mean by $\mathsf{H}(X)$ the singular homology if X is a manifold or a subset of $\mathbb{R}^d$, and simplicial homology if X is a simplicial complex. Both homologies are assumed to be defined with $\mathbb{Z}_2$ coefficients. We make similar assumptions to denote the relative homology groups $\mathsf{H}(X, A)$ for $A \subseteq X$. Notice that both $\mathsf{H}(X)$ and $\mathsf{H}(X, A)$ are vector spaces because they are defined with $\mathbb{Z}_2$ coefficients. The following known result turns out to be useful.

Proposition 1 ([3]) *Let* $\mathsf{H}(A) \to \mathsf{H}(B) \to \mathsf{H}(C) \to \mathsf{H}(D) \to \mathsf{H}(E) \to \mathsf{H}(F)$ *be a sequence of homomorphisms. If* $\text{rank}(\mathsf{H}(A) \to \mathsf{H}(F)) = \text{rank}(\mathsf{H}(C) \to \mathsf{H}(D)) = k$, *then* $\text{rank}(\mathsf{H}(B) \to \mathsf{H}(E)) = k$.

Overview of approach. We are given an ε-sample $P = \{p_i\}_{i=1}^n$ of a compact smooth m-manifold M embedded in $\mathbb{R}^d$. However, the intrinsic dimension m of M is not known, and our goal is to estimate m from the point sample P. Note that for any point $z \in \mathsf{M}$, we have that $\mathsf{H}(\mathsf{M}, \mathsf{M} - z) \cong \tilde{\mathsf{H}}(\mathbf{S}^m)$, which is the reduced homology of $\mathbf{S}^m$. This means that $\text{rank}(\mathsf{H}_i(\mathsf{M}, \mathsf{M} - z)) = 1$ if and only if $i = m$. Hence, if we can compute the rank of $\mathsf{H}_i(\mathsf{M}, \mathsf{M} - z)$ for every i, then we can recover the dimension of M. This is the approach we will follow. In Section 4, we first relate $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ with the topology of the offset of the point set P. This requires us to inspect the deformation retraction from the offset to M carefully. The relation to the offset, in turns, allows us to provably recover the rank of $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ using the so-called Vietoris Rips complex, which we detail in Section 5. One key ingredient here is to use only local neighborhoods of a sample point to obtain the estimate. First, in Section 3, we derive several technical results to prepare for the development of our approach in Section 4 and 5.

3 Local Homology of M and its Offsets

Local homology $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$. In this section, we develop a few results that we use later. First, we relate the target local homology groups $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ to some other local homology which becomes useful later for connecting to the local homology of Rips complexes that are ultimately used in the algorithm. We start by quoting the following known result.

Proposition 2 ([6]) *Let $B_r(p)$ be a closed Euclidean ball so that it intersects the m-manifold M in more than one point. If $r < \rho(\mathsf{M})$, then $\mathsf{M} \cap B_r(p)$ is a closed topological m-ball.*

The next result now follows.

Proposition 3 *Let $D \subset \mathsf{M}$ be a closed topological m-ball from the m-manifold M, and $z \in \mathsf{M}$ a point contained in the interior $\mathring{D}$ of D. Then i_* is an isomorphism where*

$$\mathsf{H}(\mathsf{M}, \mathsf{M} - \mathring{D}) \xrightarrow{i_*} \mathsf{H}(\mathsf{M}, \mathsf{M} - z).$$

We can extend Proposition 3 a little further.

Proposition 4 *Let D_1 and D_2 be two closed topological balls containing z in the interior where $D_1 \subseteq D_2 \subseteq \mathsf{M}$. The inclusion-induced homomorphisms i'_* and i_* in the following sequence are isomorphisms:*

$$\mathsf{H}(\mathsf{M}, \mathsf{M} - \mathring{D}_2) \xrightarrow{i'_*} \mathsf{H}(\mathsf{M}, \mathsf{M} - \mathring{D}_1) \xrightarrow{i_*} \mathsf{H}(\mathsf{M}, \mathsf{M} - z).$$

Local homology of the offset. We wish to relate the local homology $\mathsf{H}(\mathsf{M}, \mathsf{M} - z)$ at a point z to the local homology of an α-offset of an ε-sample $P = \{p_i\}_{i=1}^n$, defined as

$$\mathbb{X}_\alpha = \cup_{i=1}^n B_\alpha(p_i)$$

which is the union of balls centered at every p_i with radius α. For this, we will need a map to connect the two spaces, which is provided by the following projection map:

$$\pi_\alpha : \mathbb{X}_\alpha \to \mathsf{M} \text{ given by } x \mapsto \text{argmin}_{z \in \mathsf{M}} d(x, z).$$

Choose $\alpha < \rho(\mathsf{M}) - \varepsilon$. Since P is an ε-sample, no point of $\mathbb{X}_\alpha$ is $\rho(\mathsf{M})$ or more away from M. This means that no point of the medial axis of M is included in $\mathbb{X}_\alpha$. Therefore, the map π is well defined. Furthermore, by a result of [14], π is a deformation retraction for appropriate choices of parameters. In fact, under this projection map, the pre-image of a point has a nice structure (star-shaped).

For convenience denote $\theta_1 = \frac{(\varepsilon+\rho) - \sqrt{\varepsilon^2+\rho^2-6\varepsilon\rho}}{2}$ and $\theta_2 = \frac{(\varepsilon+\rho) + \sqrt{\varepsilon^2+\rho^2-6\varepsilon\rho}}{2}$ and observe that $\varepsilon \le \theta_1$ and $\theta_2 \le \rho(\mathsf{M}) - \varepsilon$ for $\varepsilon, \rho > 0$. We have:

Proposition 5 *Let $0 < \varepsilon < (3 - \sqrt{8})\rho(\mathsf{M})$ and $\theta_1 \le \alpha \le \theta_2$. Let $\mathbb{A}_\alpha = \pi_\alpha^{-1}(\mathsf{N})$ where $\mathsf{N} \subseteq \mathsf{M}$ may be either an open or a closed subset. Then $\pi_\alpha : \mathbb{A}_\alpha \to \mathsf{N}$ is a retraction and N is a deformation retract of $\mathbb{A}_\alpha$.*

Based on the above observation, the map $\pi_\alpha : (\mathbb{X}_\alpha, \mathbb{A}_\alpha) \to (\mathsf{M}, \mathsf{N})$ seen as a map on the pairs provides an isomorphism at the homology level.

Proposition 6 *Let $0 < \varepsilon < (3 - \sqrt{8})\rho$ and $\theta_1 \le \alpha \le \theta_2$. The homomorphism $\pi_{\alpha*} : \mathsf{H}(\mathbb{X}_\alpha, \mathbb{A}_\alpha) \to \mathsf{H}(\mathsf{M}, \mathsf{N})$ is an isomorphism.*

Proposition 7 *Let $0 < \varepsilon < (3 - \sqrt{8})\rho$, and $\theta_1 \le \alpha < \alpha' \le \theta_2$. Let $\mathsf{N} \subset \mathsf{N}'$ be two closed (or open) sets of M, and $\mathbb{A}_\alpha = \pi_\alpha^{-1}(\mathsf{N})$ and $\mathbb{A}_{\alpha'} = \pi_{\alpha'}^{-1}(\mathsf{N}')$. Denoting by $\text{im}(\cdot)$ the image of a map, we have*

$$\text{im}\left(\mathsf{H}(\mathbb{X}_\alpha, \mathbb{A}_\alpha) \to \mathsf{H}(\mathbb{X}_{\alpha'}, \mathbb{A}_{\alpha'})\right) \cong \text{im}\left(\mathsf{H}(\mathsf{M}, \mathsf{N}) \to \mathsf{H}(\mathsf{M}, \mathsf{N}')\right).$$

4 Local Interleaving of Offsets

Let $p \in P$ be any sample point. We show how to obtain the local homology of the projected point $\pi(p)$ on M from pairs of p's local neighborhoods in $\mathbb{X}_\alpha$. The results from the previous section already allow us to relate the local homology of the projected point $\pi(p)$ with the local homology of some local neighborhoods in $\mathbb{X}_\alpha$ (which are the pre-image of some sets in M). We now use interleaving to relate them further to local neighborhoods that are intersection of $\mathbb{X}_\alpha$ with Euclidean balls. Since $\pi(p)$ plays an important role here, we use a special symbol $\bar{p} = \pi(p)$ for it. For convenience, we introduce notations (see Figure 1):

$$\begin{aligned} \mathbb{M}_{\alpha,\beta} &= \pi_\alpha^{-1}(\mathring{B}_\beta(p) \cap \mathsf{M}),\ \mathbb{M}^{\alpha,\beta} = \mathbb{X}_\alpha - \mathbb{M}_{\alpha,\beta}, \text{ and} \\ \mathbb{B}_{\alpha,\beta} &= \mathring{B}_\beta(p) \cap \mathbb{X}_\alpha,\ \mathbb{B}^{\alpha,\beta} = \mathbb{X}_\alpha - \mathbb{B}_{\alpha,\beta}. \end{aligned}$$

The following simple observation follows from Propositions 3, 2, and 5.

Proposition 8 *Let $D_\beta = B_\beta(p) \cap \mathsf{M}$. For $0 < \varepsilon < (3 - \sqrt{8})\rho$, $\varepsilon < \beta < \rho(\mathsf{M})$ and $\theta_1 \le \alpha \le \theta_2$, the maps $\pi_{\alpha*}$ and i_* are isomorphisms in the sequence: $\mathsf{H}(\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\beta}) \xrightarrow{\pi_{\alpha*}} \mathsf{H}(\mathsf{M}, \mathsf{M} - \mathring{D}_\beta) \xrightarrow{i_*} \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p})$.*

Now set $\delta = \alpha + 3\varepsilon$. Consider any $z \in \mathsf{M}$. Since any point $x \in \pi_\alpha^{-1}(z)$ resides within a ball $B_\alpha(p_i)$ for some $p_i \in P$, we have that

$$\begin{aligned} d(x, z) &= d(x, \pi(x)) && (1) \\ &\le d(x, \pi(p_i)) \le d(x, p_i) + d(p_i, \pi(p_i)) && (2) \\ &\le \alpha + \varepsilon = \delta - 2\varepsilon. && (3) \end{aligned}$$

It follows that for any $\lambda \in (\varepsilon, \rho(\mathsf{M}) - \delta)$ we get the following inclusions.

$$\mathbb{M}_{\alpha,\lambda} \subset \mathbb{B}_{\alpha,\lambda+\delta} \subset \mathbb{M}_{\alpha,\lambda+2\delta} \subset \mathbb{B}_{\alpha,\lambda+3\delta} \subset \mathbb{M}_{\alpha,\lambda+4\delta}.$$

Taking the complements, a new filtration in the reverse direction is generated:

$$\mathbb{M}^{\alpha,\lambda+4\delta} \subset \mathbb{B}^{\alpha,\lambda+3\delta} \subset \mathbb{M}^{\alpha,\lambda+2\delta} \subset \mathbb{B}^{\alpha,\lambda+\delta} \subset \mathbb{M}^{\alpha,\lambda}.$$

Considering each space as a topological pair, the nested sequence becomes

$$\begin{aligned} (\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\lambda+4\delta}) &\subset (\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+3\delta}) \\ &\subset (\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\lambda+2\delta}) \\ &\subset (\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+\delta}) \\ &\subset (\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\lambda}) \qquad (4) \end{aligned}$$

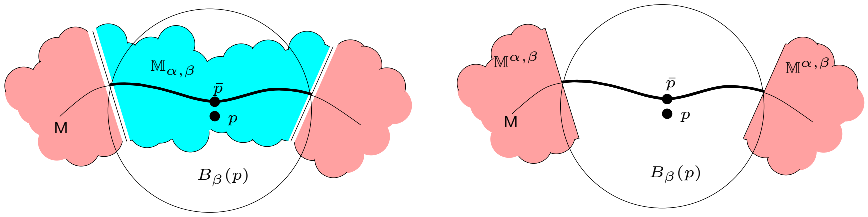

$\mathbb{M}_{\alpha,\beta}$
$\bar{p}$
p
M
$B_\beta(p)$
$\mathbb{M}^{\alpha,\beta}$
$\mathbb{M}^{\alpha,\beta}$
$\bar{p}$
p
M
$B_\beta(p)$

Figure 1: The spaces $\mathbb{M}_{\alpha,\beta}$ shown in cyan (left) and $\mathbb{M}^{\alpha,\beta}$ shown in pink (right).

Inclusion between topological pairs induces a homomorphism between their relative homology groups. Therefore, the following relative homology sequence holds.

$$\begin{aligned} \mathsf{H}(\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\lambda+4\delta}) &\to \mathsf{H}(\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+3\delta}) \\ &\to \mathsf{H}(\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\lambda+2\delta}) \\ &\to \mathsf{H}(\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+\delta}) \\ &\to \mathsf{H}(\mathbb{X}_\alpha, \mathbb{M}^{\alpha,\lambda}) \qquad (5) \end{aligned}$$

Let $\epsilon \leq \alpha' \leq \rho(\mathsf{M}) - \epsilon$ and $\delta' = \alpha' + 3\varepsilon$. Similar to sequence (4), for any $\lambda' \in (\varepsilon, \rho(\mathbb{M}) - 4\delta')$ we have:

$$\begin{aligned} (\mathbb{X}_{\alpha'}, \mathbb{M}^{\alpha',\lambda'+4\delta'}) &\subset (\mathbb{X}_{\alpha'}, \mathbb{B}^{\alpha',\lambda'+3\delta'}) \\ &\subset (\mathbb{X}_{\alpha'}, \mathbb{M}^{\alpha',\lambda'+2\delta'}) \\ &\subset (\mathbb{X}_{\alpha'}, \mathbb{B}^{\alpha',\lambda'+\delta'}) \\ &\subset (\mathbb{X}_{\alpha'}, \mathbb{M}^{\alpha',\lambda'}) \qquad (6) \end{aligned}$$

The stated range of λ, λ' is valid if $\alpha, \alpha' < \frac{\rho(\mathsf{M})-13\varepsilon}{4}$. We also need $\theta_1 \leq \alpha, \alpha'$. These two conditions are satisfied for $\varepsilon < \frac{\rho(\mathsf{M})}{22}$. Let $\theta_2' = \frac{\rho(\mathsf{M})-13\varepsilon}{4}$.

Proposition 9 *Let $0 < \varepsilon < \frac{\rho(\mathsf{M})}{22}$, and $\theta_1 \leq \alpha \leq \alpha' \leq \theta_2'$. Set $\delta = \alpha + 3\varepsilon$ and $\delta' = \alpha' + 3\varepsilon$. For $\varepsilon < \lambda' < \rho(\mathsf{M}) - 4\delta'$ and $\lambda \geq \lambda' + 2(\alpha' - \alpha)$, we have,*

$$\operatorname{im} \left(\mathsf{H}(\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+3\delta}) \to \mathsf{H}(\mathbb{X}_{\alpha'}, \mathbb{B}^{\alpha',\lambda'+\delta'}) \right) \cong \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p}). \qquad (7)$$

In particular,

$$\operatorname{im} \left(\mathsf{H}(\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+3\delta}) \to \mathsf{H}(\mathbb{X}_\alpha, \mathbb{B}^{\alpha,\lambda+\delta}) \right) \cong \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p}).$$

Finally, we intersect each set with a sufficiently large ball $B_r(p)$ so that we only need to inspect within the neighborhood $B_r(p)$ of p. Specifically, denote $\mathbb{X}_{\alpha,r} = \mathbb{X}_\alpha \cap B_r(p)$ and $\mathbb{X}_{\alpha,r}^\beta = \mathbb{X}_{\alpha,r} \cap B^\beta(p)$. We obtain the next proposition by applying the Excision theorem (details in the full version [7]).

Proposition 10 *Let all the parameters satisfy the same conditions as in Proposition 9. Then, for $r > \lambda + 5\delta$, we have:*

$$\operatorname{im} \left(\mathsf{H}(\mathbb{X}_{\alpha,r}, \mathbb{X}_{\alpha,r}^{\lambda+3\delta}) \to \mathsf{H}(\mathbb{X}_{\alpha',r}, \mathbb{X}_{\alpha',r}^{\lambda'+\delta'}) \right) \cong \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p}).$$

In particular,

$$\operatorname{im} \left(\mathsf{H}(\mathbb{X}_{\alpha,r}, \mathbb{X}_{\alpha,r}^{\lambda+3\delta}) \to \mathsf{H}(\mathbb{X}_{\alpha,r}, \mathbb{X}_{\alpha,r}^{\lambda+\delta}) \right) \cong \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p}).$$

In fact, one can relax the parameters, and the image homology $\operatorname{im} \left(\mathsf{H}(\mathbb{X}_{\alpha,r}, \mathbb{X}_{\alpha,r}^{\beta_2}) \to \mathsf{H}(\mathbb{X}_{\alpha',r}, \mathbb{X}_{\alpha',r}^{\beta_1}) \right)$ captures (that is, is isomorphic to) the local homology $\mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p})$ as long as $\beta_1 \geq \alpha' + 4\varepsilon$, $\beta_2 \geq \beta_1 + \alpha + \alpha' + 6\varepsilon$ and $r > \beta_2 + 2\alpha + 6\varepsilon$.

5 Interleaving Nerves and Rips complexes

We now relate the relative homology of pairs as in Proposition 10 to the relative homology of pairs in Rips complexes so that they can computed. Our algorithm works on these pairs of Rips complexes to derive the local homology at a point on M. As before, let $p \in P$ be a point from the sample.

Nerves of spaces. Consider the space $\mathbb{X}_{\alpha,r} = \mathbb{X}_\alpha \cap B_r(p)$. The connection of such spaces with simplicial complexes (Vietoris-Rips complex in particular) is made through the so-called nerve of a cover. In general, let $\mathcal{U}$ be a finite collection of sets. The *nerve $\mathcal{N}\mathcal{U}$ of $\mathcal{U}$* is a simplicial complex whose simplices are given by all subsets of $\mathcal{U}$ whose members have a non-empty common intersection. That is,

$$\mathcal{N}\mathcal{U} := \{\mathcal{A} \subseteq \mathcal{U} \mid \cap\mathcal{A} \neq \emptyset\}.$$

The set $\mathcal{U}$ forms a *good cover* of the union $\bigcup \mathcal{U}$ if the intersection of any subsets of $\mathcal{U}$ is either empty or contractible. The Nerve Lemma states that if $\mathcal{U}$ is a good cover of $\bigcup \mathcal{U}$, then $\mathcal{N}\mathcal{U}$ is homotopic to $\bigcup \mathcal{U}$, denoted by $\mathcal{N}\mathcal{U} \approx \bigcup \mathcal{U}$.

Now consider the set of sets $\mathcal{X}_{\alpha,r} = \{B_\alpha(p_i) \cap B_r(p) \mid p_i \in P\}$; note that $\mathbb{X}_{\alpha,r} = \bigcup \mathcal{X}_{\alpha,r}$. Since each set in $\mathcal{X}_{\alpha,r}$ is convex, $\mathcal{X}_{\alpha,r}$ forms a good cover of $\mathbb{X}_{\alpha,r}$ and thus $\mathcal{N}\mathcal{X}_{\alpha,r} \approx \mathbb{X}_{\alpha,r}$ by the Nerve Lemma. Furthermore, it follows from Lemma A.5 of [15] that for $r > \beta + 2\alpha$, the set $\mathcal{X}_{\alpha,r}^\beta = \{B_\alpha(p_i) \cap B_r(p) \cap B^\beta(p)\}_{i \in [1,n]}$ also form a good cover of $\bigcup \mathcal{X}_{\alpha,r}^\beta (= \mathbb{X}_{\alpha,r}^\beta)$. Thus, we have $\mathcal{N}\mathcal{X}_{\alpha,r}^\beta \approx \mathbb{X}_{\alpha,r}^\beta$. We can now convert the relative homology between $\mathbb{X}_{\alpha,r}$ and $\mathbb{X}_{\alpha,r}^\beta$ to the homology of their nerves. In particular, we have the following result. The proof relies heavily on the proof of Lemma 3.4 of [3] which gives a crucial commutative result for the space and its nerve.

Lemma 11 *Let all the parameters satisfy the same conditions as in Proposition 9. Then, for $r > \lambda + 5\delta$:*

$$\mathrm{im}\left(\mathsf{H}(\mathcal{N}\mathcal{X}_{\alpha,r}, \mathcal{N}\mathcal{X}_{\alpha,r}^{\lambda+3\delta}) \to \mathsf{H}(\mathcal{N}\mathcal{X}_{\alpha',r}, \mathcal{N}\mathcal{X}_{\alpha',r}^{\lambda'+\delta'})\right) \cong \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p}).$$

Relating nerves and Rips complexes. First, we recall that for $\alpha \geq 0$, the *Čech complex* $C^\alpha(Q)$ of a point set Q is the nerve of the cover $\{B_\alpha(q_i) : q_i \in Q\}$ of $\cup B_\alpha(q_i) = \mathbb{X}_\alpha$. The *Vietoris-Rips* (Rips in short) complex $\mathcal{R}^\alpha(Q)$ is the maximal complex induced by the edge set $\{(p_j, p_k) \mid d(p_j, p_k) \leq \alpha\}$. It is well known that for any point set Q, the following holds:

$$C^\alpha(Q) \subset \mathcal{R}^{2\alpha}(Q) \subset C^{2\alpha}(Q).$$

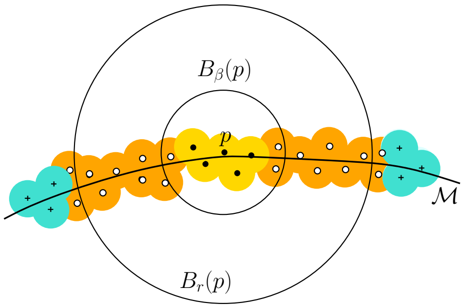
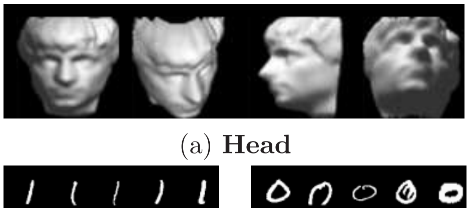


Figure 2: The points in $P_{\alpha,r}$ are the centers of yellow and brown balls whereas the centers of the brown balls are in $P^\beta_{\alpha,r}$.

Define $P_{\alpha,r} = \{p_i \in P \mid B_\alpha(p_i) \cap B_r(p) \neq \emptyset\}$. Obviously, $P_{\alpha,r}$ forms the vertex set for the nerve $\mathcal{N}\mathcal{X}_{\alpha,r}$. Similarly, let $P^\beta_{\alpha,r} = \{p_i \in P_{\alpha,r} \mid B_\alpha(p_i) \cap B^\beta(p) \neq \emptyset\}$ denote the vertex set of $\mathcal{N}\mathcal{X}^\beta_{\alpha,r}$. See Figure 2 for an example, where the union of solid and empty dots forms the set of points $P_{\alpha,r}$, while $P^\beta_{\alpha,r}$ consists the set of empty dots. Note that from the definition, it follows that $P^\beta_{\alpha,r} \subset P_{\alpha,r}$ and $P^\beta_{\alpha,r} \subset P^{\beta'}_{\alpha,r}$ for $\beta' < \beta$. Furthermore, as the offset $\mathbb{X}_\alpha$ grows, it is immediate that $P_{\alpha,r} \subset P_{\alpha',r}$ and $P^\beta_{\alpha,r} \subset P^\beta_{\alpha',r}$ for $\alpha < \alpha'$.

Each element in the good cover $\mathcal{X}_{\alpha,r}$ or $\mathcal{X}^\beta_{\alpha,r}$ is in the form of $B_\alpha(p_i) \cap B_r(p)$ or $B_\alpha(p_i) \cap B_r(p) \cap B^\beta(p)$. Since the Čech complex of a point set is the nerve of the set of balls centered at these points, it follows easily that

$$\mathcal{N}\mathcal{X}_{\alpha,r} \subset C^\alpha(P_{\alpha,r}) \subset \mathcal{R}^{2\alpha}(P_{\alpha,r}) \quad \text{and} \quad \mathcal{N}\mathcal{X}^\beta_{\alpha,r} \subset C^\alpha(P^\beta_{\alpha,r}) \subset \mathcal{R}^{2\alpha}(P^\beta_{\alpha,r}). \tag{8}$$



(a) **Head**

(b) **D1**

(c) **D0**

Figure 3: Image data : **Head**, **D1** and **D0**

Claim 12 *(i)* $\mathcal{R}^{2\alpha}(P_{\alpha,r}) \subset \mathcal{N}\mathcal{X}_{3\alpha,r}$, *and (ii)* $\mathcal{R}^{2\alpha}(P^\beta_{\alpha,r}) \subset \mathcal{N}\mathcal{X}^\beta_{3\alpha,r}$.

Combining the above claim and Eqn (8), we get an interleaving sequence between the nerve and the Rips complexes (see full version [7] for details) from which we can derive our main result:

Theorem 13 *Let* $0 < \varepsilon < \frac{\rho(\mathsf{M})}{58}$ *and* $\theta_1 \leq \alpha \leq \frac{\rho(\mathsf{M})-13\varepsilon}{22}$. *Furthermore, let* η_1 *and* η_2 *be such that* $\varepsilon < \eta_1, \eta_2 < \rho(\mathsf{M})$, $\eta_1 \geq 9\alpha + 4\varepsilon$, *and* $\eta_2 \geq \eta_1 + 12\alpha + 6\varepsilon$. *The inclusion*

$$j_\alpha : (\mathcal{R}^{2\alpha}(P_{\alpha,r}), \mathcal{R}^{2\alpha}(P^{\eta_2}_{\alpha,r})) \hookrightarrow (\mathcal{R}^{6\alpha}(P_{3\alpha,r}), \mathcal{R}^{6\alpha}(P^{\eta_1}_{3\alpha,r}))$$

satisfies $\mathrm{im}(j_{\alpha *}) \cong \mathsf{H}(\mathsf{M}, \mathsf{M} - \bar{p})$ *for any* $r \geq \eta_1 + \eta_2$.

Algorithm. Given a sample point $p = p_i$, our algorithm first constructs the necessary Rips complexes as specified in Theorem 13 for some parameters $\alpha < \eta_1 < \eta_2 < r$. For simplicity, rewrite $j_\alpha : (A_1, B_1) \hookrightarrow (A_2, B_2)$ where $B_1 \subset A_1 \subset A_2$ and $B_1 \subset B_2 \subset A_2$. After obtaining the necessary Rips complexes, one possible method for computing $\mathrm{im}(j_{\alpha *})$ would be to cone the subcomplexes B_1 and B_2 with a dummy vertex w to obtain an inclusion $\iota : A_1 \cup (w * B_1) \hookrightarrow A_2 \cup (w * B_2)$ where $w * B_j = B_j \cup \{w * \sigma | \sigma \in B_j\}$ is the cone on B_j $(j = 1, 2)$. It is easy to see that $\mathrm{im}(j_{\alpha *}) \cong \mathrm{im}(\iota_*)$. Then, the standard persistent homology algorithm can be applied. However, the cone operations may add many unnecessary simplices slowing down the computation. Instead, we order the simplices in A_2 properly to build a filtration so that the rank of $\mathrm{im}(j_{\alpha *})$ can be read off from the reduced boundary matrix built from the filtration. The details of this algorithm can be found in the full version [7].

6 Experimental results

Due to the lack of space, we refer the readers to the full version [7] for the implementation details: in particular to see how we choose points to perform our dimension estimation algorithm and how we filter false positives. The full version also includes experimental results of our algorithm on synthetic data sets. In this abstract, we only present the comparison results of our algorithm with several state-of-the-art algorithms on real data; which is shown in Table 1.

Specifically, the real data contains images of a 2D translation of a smaller image within a black image (**Shift**) (see [4]), a rotating head (**Head**, Fig. 3(a)), handwritten 1's (**D1**, Fig. 3(b)) and 0's (**D0**, Fig. 3(c)) from MNIST database. Our method is compared with the dimension detection method via slivers (SLIVER) [4], the maximum likelihood estimation (MLE) [12],

	Shift	**Head**	**D1**	**D0**
Ours	2	3	4	3
SLIVER	3	4	3	2
MLE	4.27	4.31	11.47	14.86
MA	3.35	4.47	10.77	13.93
PN	3.62	3.98	6.22	8.86
LPCA	3	3	5	8.86
ISOMAP	2	3	5	$[3, 6]$

Table 1: Comparison results

the manifold adaptive method (MA) [9], the packing number method (PN) [11], the local PCA (LPCA) [5], and the isomap method (ISOMAP) [16]. Notice that although **Shift** is uniform and noise free, only ISOMAP and ours get the correct dimension. The dimension of **Head** is considered to be around 3 or 4 in the literature. Ours falls into this range. Although the ground truth dimensions for **D1** and **D0** are unknown, ours along with SLIVER, PN, LPCA and ISOMAP report dimensions in range $[3, 7]$ for **D1** and in range $[2, 9]$ for **D0**.

7 Conclusions

In this paper, we present a topological method to estimate the dimension of a manifold from its point samples with a theoretical guarantee. The use of local topological structures helps to alleviate the dependency of our method on the regularity of point samples, and the use of persistent homology for a pair of homology groups (instead of a single homology group) helps to increase its robustness.

It will be interesting to investigate other data analysis problems where topological methods, especially those based on local topological information (yields to efficient computations), may be useful. Currently, we have conducted some preliminary experiments to demonstrate the performance of our algorithm. It will be interesting to conduct large-scale experiments under a broad range of practical scenarios, so as to better understand data in those contexts.

Acknowledgment

We acknowledge the support of the NSF grants CCF 1318595, CCF 1319406, CCF 1116258 that helped conducting this research.

References

[1] P. Bendich, D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov. Inferring local homology from sampled stratified spaces. In *Proc. 48th Ann. IEEE Sympos. Foundat. Comp. Sci.*, pages 536–546, 2007.

[2] P. Bendich, B. Wang, and S. Mukherjee. Local homology transfer and stratification learning. In *Proc. 23rd Ann. ACM-SIAM Sympos. Discrete Alg.*, pages 1355–1370, 2012.

[3] F. Chazal and S. Oudot. Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, SCG '08, pages 232–241, 2008.

[4] S.-W. Cheng and M.-K. Chiu. Dimension detection via slivers. In *Proc. 20th Ann. ACM-SIAM Sympos. Discrete Alg.*, pages 1001–1010, 2009.

[5] S.-W. Cheng, Y. Wang, and Z. Wu. Provable dimension detection using principal component analysis. In *Proc. 21st Ann. Sympos. Comput. Geom.*, pages 208–217, 2005.

[6] T. K. Dey. *Curve and surface reconstruction: Algorithms with mathematical analysis*. Cambridge University Press, New York, 2006.

[7] T. K. Dey, F. Fan, and Y. Wang. Dimension detection with local homology. available from authors' web-page.

[8] T. K. Dey, J. Giesen, S. Goswami, and W. Zhao. Shape dimension and approximation from samples. *Discrete Comput. Geom.*, 29:419–434, 2003.

[9] A. M. Farahmand, C. Szepesvári, and J.-Y. Audibert. Manifold-adaptive dimension estimation. In *Proc. 24th Conf. Machine Learning (ICML).*, pages 265–272, 2007.

[10] J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds with high co-dimension. In *Proc. 19th Ann. Sympos. Comput. Geom.*, pages 329–337, 2003.

[11] B. Kégl. Intrinsic dimension estimation using packing numbers. In *Neural Infor. Proc. Sys. Foundation (NIPS)*, pages 681–688, 2002.

[12] E. levina and P. J. bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17*, pages 777–784, 2005.

[13] A. V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for estimating intrinsic dimension. In *Proc. SampTA*, 2011.

[14] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39:419–441, March 2008.

[15] P. Skraba and B. Wang. Approximating local homology from samples. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 174–192, 2014.

[16] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.