

A TOPOLOGICAL VIEW OF UNSUPERVISED LEARNING FROM NOISY DATA*

P. NIYOGI[†], S. SMALE[‡], AND S. WEINBERGER[§]

Abstract. In this paper, we take a topological view of unsupervised learning. From this point of view, clustering may be interpreted as trying to find the number of connected components of any underlying geometrically structured probability distribution in a certain sense that we will make precise. We construct a geometrically structured probability distribution that seems appropriate for modeling data in very high dimensions. A special case of our construction is the mixture of Gaussians where there is Gaussian noise concentrated around a finite set of points (the means). More generally we consider Gaussian noise concentrated around a low dimensional manifold and discuss how to recover the homology of this underlying geometric core from data that do not lie on it. We show that if the variance of the Gaussian noise is small in a certain sense, then the homology can be learned with high confidence by an algorithm that has a weak (linear) dependence on the ambient dimension. Our algorithm has a natural interpretation as a spectral learning algorithm using a combinatorial Laplacian of a suitable data-derived simplicial complex.

Key words. topology, data, manifolds

AMS subject classifications. Primary, 62-07; Secondary, 52B55, 57N65, 58-04, 62G99, 65Y20

DOI. 10.1137/090762932

1. Introduction. An unusual and arguably ubiquitous characteristic of modern data analysis is the high dimensionality of the data points. One can think of many examples from image processing and computer vision, acoustics and signal processing, bioinformatics, neuroscience, finance, and so on where this is the case. The strong intuition of researchers has always been that naturally occurring data cannot possibly “fill up” the high dimensional space uniformly, rather it must concentrate around lower dimensional structures. A goal of exploratory data analysis or unsupervised learning is to extract this kind of low dimensional structure with the hope that this will facilitate further processing or interpretation.

For example, principal components analysis is a widely used methodological tool to project the high dimensional data linearly into a lower dimensional subspace along the directions of maximal variation in a certain sense. This serves the role of smoothing the data and reducing its essential dimensions before further processing. Another canonical unsupervised technique is clustering, which has also received considerable attention in statistics and computer science. In this paper, we wish to develop the point of view that clustering is a kind of topological question one is asking about the data and the probability distribution underlying it: in some sense one is trying to partition the underlying space into some natural *connected components*. Following this line of thinking leads one to ask whether more general topological properties may

*Received by the editors June 23, 2009; accepted for publication (in revised form) March 7, 2011; published electronically June 2, 2011.

<http://www.siam.org/journals/sicomp/40-3/76293.html>

[†]This author is deceased. Former address: Department of Computer Science, University of Chicago, 1100 E. 58th Street, Ryerson 167, Chicago, IL 60637.

[‡]Department of Mathematics, City University of Hong Kong, Tat Chee Ave., Hong Kong (smale@math.berkeley.edu). The work of this author was partially supported by a joint NSF grant.

[§]Department of Mathematics, University of Chicago, 1100 E. 58th Street, Eckhart Hall, Chicago, IL 60637 (shmuel@math.uchicago.edu). The work of this author was partially supported by NSF grant 0805913, NSF-SGER grant DMS-0852227, and STOMP, a DARPA program.

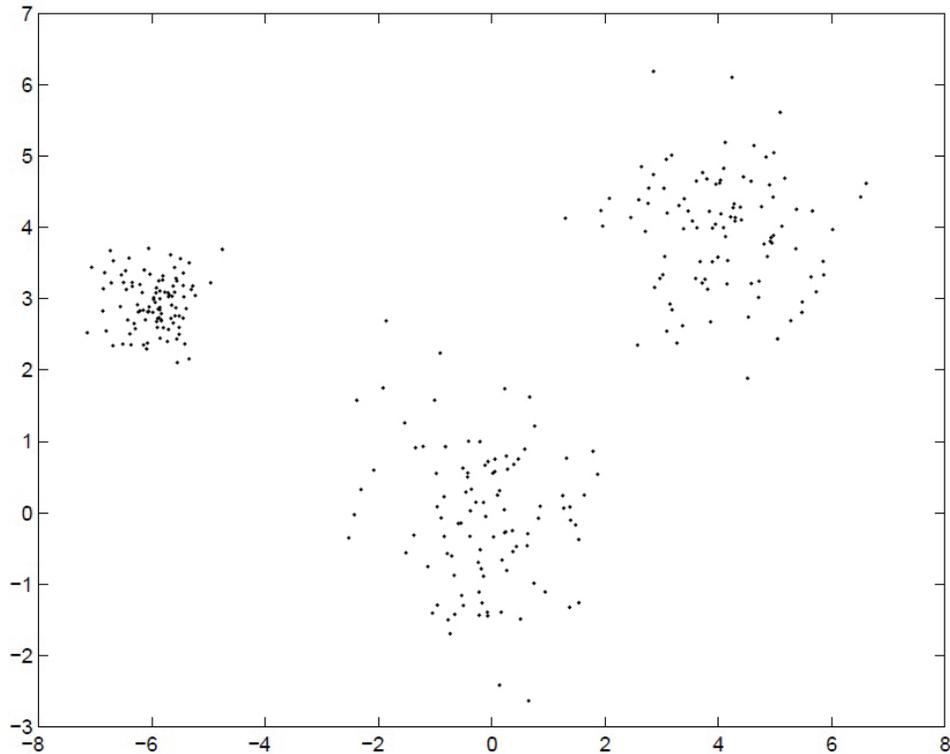


FIG. 1. A random data set that is consistent with a mixture of Gaussians.

be inferred from data. As we shall see, from this the homology learning question follows naturally.

As a first example, consider Figure 1, which consists of a cloud of points in \mathbb{R}^2 . The viewer immediately sees three clusters of points. This picture motivates a conceptualization of clustering as data arising from a mixture of distributions, each of which may be suitably modeled as a Gaussian distribution around its centroid. This is a fairly classical view of clustering that has received a lot of attention in statistics over the years and more recently in computer science as well.

In contrast, consider Figure 2. Here one sees three clusters again. But these are hardly like Gaussian blobs! In fact, one notices immediately that two of the clusters are like circles while one is like a Gaussian blob. This picture motivates a different conceptualization of clustering as trying to find the connected components of the data set at hand—this has led to the recent surge of interest in spectral clustering and related algorithms (see [10] and the references therein).

Now if one were interested in simply learning the “number of clusters,” a natural spectral algorithm would proceed by building a suitable nearest neighbor graph with vertices identified with data points, connecting nearby data points to each other, and finding the number of connected components in this data-derived graph. But if one wanted to learn further structure, then one needs to do more. Building on the notion that the number of connected components is related to the zeroth homology and is one of the simplest topological invariants of the space, we see that it is natural to ask if one could learn higher order homologies as well. More generally, one may ask the

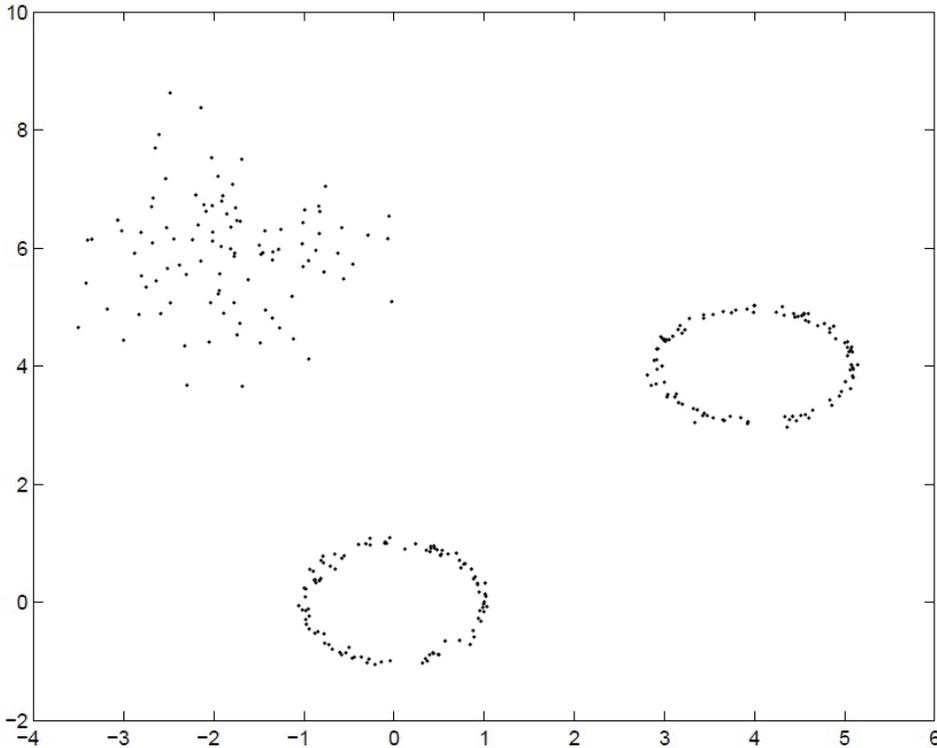


FIG. 2. A random data set in \mathbb{R}^2 that is not obviously consistent with a mixture of a small number of Gaussians. Yet it seems to the viewer that there are clearly three groups of data.

following questions:

1. What are flexible, nonparametric models of probability distributions in high dimensional spaces?
2. What sorts of structural information about these distributions can be estimated from random data? In particular, can one avoid the *curse of dimensionality* in the associated inference problems?

In this paper, we explore these two questions in a certain setting. We follow the intuition that in high dimensional spaces the underlying probability distribution is far from uniform and must in fact concentrate around lower dimensional structures. These lower dimensional structures need not be linear, and so as a first step we consider them to be submanifolds of the ambient space. The data then concentrate around this submanifold \mathcal{M} though it does not lie exactly on it. This allows us to define a family of geometrically structured probability distributions in a high dimensional space where the distribution ρ has support *on all of* \mathbb{R}^D though it concentrates around a low dimensional submanifold. This includes as a special case the mixture of Gaussians, a classical and much studied family of probability distributions. We introduce this geometrically structured family in the next section. We next consider the task of estimating the homology of the underlying manifold from noisy data obtained from the geometrically structured probability distribution concentrated around this manifold. Our main result is that a two stage variant of the algorithmic framework proposed in Niyogi, Smale, and Weinberger [17] is robust to Gaussian noise, provided the noise is

low. These algorithms may be interpreted as a kind of generalized spectral clustering using the combinatorial Laplacian of a suitable data-derived simplicial complex. In particular, the sample complexity of the algorithm depends exponentially on the dimension of the manifold but depends very weakly on the ambient dimension in this setting. In this sense our results are analogous to the findings of [9] and later works (see [1, 19], among others) which show that polynomial time algorithms for estimating mixtures of Gaussians may be obtained, provided the variance of the Gaussians in question is small in relation to the distance between their centers.¹ It is impossible to do justice to the abundant literature on this subject and its complicated history here. We mention in particular the work of Amenta and Bern [1] on surface recognition and its subsequent simplifications and extensions to higher dimensions [2, 5, 6, 7]; the development of persistent homology—see the useful survey by Edelsbrunner and Harer [11] and also [8, 12, 16, 20]; the use of Laplacian methods for learning [3, 4]; and our own work on homotopical inference [17] that this work extends.

2. Problem formulation and results. In this section we describe a geometrically structured model of a probability distribution in a high dimensional space. We then describe our main result that asserts that it is possible to learn structural aspects of this probability distribution without encountering the curse of dimensionality.

2.1. Models of probability distribution and noise. The manifold \mathcal{M} is conceptualized as a platonic ideal: the geometric core of a probability distribution centered on it. Data are drawn from this distribution, and thus we receive a *noisy, point cloud* in a high dimensional space. We formalize this as follows.

Let \mathcal{M} be a compact, smooth submanifold of \mathbb{R}^N without boundary. For any $p \in \mathcal{M}$, denote the tangent space at p by T_p and the normal space by $N_p = T_p^\perp$. Since \mathcal{M} is a submanifold, we have $p \in \mathcal{M} \subset \mathbb{R}^N$, and T_p and T_p^\perp may be identified with affine subspaces of dimension d and $N - d$, respectively. With this identification there are canonical maps (respectively) from the tangent bundle $T\mathcal{M}$ and the normal bundle $N\mathcal{M}$ to \mathbb{R}^N . Now consider a probability density function P on $N\mathcal{M}$. Then for any $(x, y) \in N\mathcal{M}$ (where $x \in \mathcal{M}$ and $y \in T_x^\perp$)

$$P(x, y) = P(x)P(y|x).$$

The marginal $P(x)$ is supported intrinsically on the manifold \mathcal{M} (the reader can think of $P(x)$ as being the intrinsic volume associated to the Riemannian structure on \mathcal{M} inherited from Euclidean space), while the conditional $P(y|x)$ is the noise in the normal direction. This probability distribution can be pushed down to \mathbb{R}^N by the canonical map from $N\mathcal{M}$ to \mathbb{R}^N . This is the probability distribution defined on \mathbb{R}^N according to which data are assumed to be drawn.

Remark. A competing noise model could be considered where the noise is defined directly on the Euclidean space and is (something like) the convolution of a measure on the manifold with a normal distribution. Such measures can be handled by the same techniques as the ones used here and, indeed, present fewer technical problems. (See the remark after Theorem 2.) The model we have chosen is reasonable if one imagines that there is already noise in “the platonic world” where the data are assumed to be sampled from a geometric structure. Then it seems that noise should be thought of as lying in a vector bundle over the geometric structure. This is then mapped into

¹In the case of mixtures of Gaussians, substantial progress has been made since the algorithmic insights of [9] so that the requirements on the noise have been weakened.

a Euclidean space by some function, and that provides the coordinate values of the data point that we directly apprehend.

One may ask whether the homology of \mathcal{M} can be inferred from examples drawn according to P and what the complexity of this inference problem is. We investigate this under the *strong variance condition*. This amounts to two separate assumptions:

1. $0 < a \leq P(x) \leq b$ for all $x \in \mathcal{M}$.
2. $P(y|x)$ is normally distributed with mean 0 and covariance matrix $\sigma^2 \mathbf{I}$, where \mathbf{I} is the $(D - d) \times (D - d)$ identity matrix.

2.1.1. Mixture of Gaussians. The most obvious special case of our framework is the mixture of Gaussians. Consider a probability distribution P on \mathbb{R}^N given by

$$P(x) = \sum_{i=1}^k w_i \mathcal{N}(x; \mu_i, \Sigma_i),$$

where $\mathcal{N}(x; \mu_i, \Sigma_i)$ is the density function for the normal distribution with mean μ_i and covariance matrix Σ_i . The weights of the mixture $w_i > 0$ sum to 1, i.e., $\sum_{i=1}^k w_i = 1$. This is a standard workhorse for density modeling and is widely used in a variety of applications.

This probability distribution may be related to our setting as follows. Consider a (zero-dimensional) manifold consisting simply of k points (identified with $\mu_1, \mu_2, \dots, \mu_k$ in \mathbb{R}^N). Thus $\mathcal{M} = \{\mu_1, \mu_2, \dots, \mu_k\}$. Let P be a probability distribution on \mathcal{M} given by $P(\mu_i) = w_i$. This manifold consists of k connected components and the normal fiber N_x for each $x \in \mathcal{M}$ has codimension D . This N_x is isomorphic to the Euclidean space \mathbb{R}^D where the origin is identified with μ_i . The probability distribution $P(y|x)$ (where $y \in N_x$) is modeled as a single Gaussian with mean 0 and variance Σ_i .

Thus if one is given a collection x_1, \dots, x_n of n points sampled from a mixture of Gaussians, one may conceptualize the data as noisy (Gaussian noise) samples from an underlying manifold. If one were able to learn the homology of the underlying manifold from such noisy samples, then one would realize (through the zeroth Betti number) that the underlying manifold has k connected components, which is equal to the number of Gaussians and the number of clusters. One would also realize (through the higher Betti numbers) that each such connected component retracts to a point. Thus, one would be able to distinguish the case of Figure 2 from Figure 1 automatically.

2.2. Main theorem. More generally, in our case, \mathcal{M} is a well-conditioned d -dimensional submanifold of \mathbb{R}^D . The conditioning of the manifold is characterized by the quantity τ that is also used in the statements of our theorems in [17]. τ is defined as the largest number having the property: The open normal bundle about \mathcal{M} of radius r is imbedded in \mathbb{R}^N for every $r < \tau$. Its image Tub_τ is a tubular neighborhood of \mathcal{M} with its canonical projection map

$$\pi_0 : \text{Tub}_\tau \rightarrow \mathcal{M}.$$

τ is the feature size introduced in [1] and that arises in many other investigations. Note that τ encodes both local curvature considerations as well as global ones: If \mathcal{M} is a union of several components, then τ bounds their separation. For example, if \mathcal{M} is a sphere, then τ is equal to its radius. If \mathcal{M} is an annulus, then τ is the separation of its components or the smaller radius. Our main theorem may be summarized as follows.

THEOREM 1 (main theorem). *Consider a probability distribution P satisfying the strong variance condition described previously in section 2.1. Let $\delta > 0$ be given; then as long as the variance σ^2 satisfies the bound*

$$(1) \quad \sqrt{8(D-d)}\sigma < c \frac{\sqrt{9}-\sqrt{8}}{9}\tau$$

for some $c < 1$, the algorithm described below will compute the homology of \mathcal{M} from a random sample of $f(\tau, \delta, c, d, D, \text{diam}(\mathcal{M}))$ points with probability $1-\delta$. (The function f is explicit and the diameter can be replaced with $\text{Vol}(\mathcal{M})$ in this estimate.) Further, if the codimension is high, in particular, if

$$D-d > A \left(\log \left(\frac{1}{a} \right) + Kd \log \left(\frac{1}{\tau} \right) \right)$$

for suitable constants $A, K > 0$, the sample complexity is independent of D . Therefore the only place where the ambient dimension D enters in the computational complexity of the algorithm is in reading the data points (linear in D).

Some further remarks are worthwhile.

1. It is worth emphasizing that the probability distribution P is supported on all of \mathbb{R}^D . Even so, the fact that it concentrates around a low dimensional structure (\mathcal{M}) allows one to beat the curse of dimensionality if $d \ll D$ and the noise σ is sufficiently small. A number of previous results (for example, [18, 15, 14, 17]) have shown how learning rates depend only on d if the probability distribution is supported on a d -manifold. It has been unclear from these prior works whether such a low dimensional rate would still hold if the distribution was supported on all of \mathbb{R}^D but concentrated around \mathcal{M} . Our results provide an answer to this question in the context of learning homology.
2. The condition on the noise σ may be seen as analogous to a similar condition on the mixture of Gaussians assumed in the breakthrough paper [9]. As discussed earlier, the mixture of Gaussians corresponds to the special case in which \mathcal{M} is a set of k points (say $\mu_1, \dots, \mu_k \in \mathbb{R}^D$). In that case, τ is simply given by

$$\tau = \frac{1}{2} \min_{i,j} \|\mu_i - \mu_j\|.$$

So the strong variance condition amounts to stipulating that the variance of the Gaussians is small relative to the distance between their means in a manner that is similar in spirit to the assumption of [9].

3. One complication that we potentially have to deal with is a feature of our more general class of probability distributions and does not arise in the case of a mixture of a finite number of Gaussians. There can be points $x \in \mathbb{R}^N, x \notin \mathcal{M}$ off the manifold where the density function blows up; i.e., the measure of a sufficiently small ball around x is very large compared to the measure of similarly sized balls around points on the manifold. As an example of such *hotspots*, consider the simple case of a circle S^1 embedded in \mathbb{R}^2 in the standard way. Let $P(x)$ (for $x \in \mathcal{M}$) be the uniform density on the circle and let $P(y|x)$ be $\mathcal{N}(0, \sigma^2)$ be the one-dimensional Gaussian along the normal fibers. Now consider the measure μ induced on \mathbb{R}^2 by this geometrically structured distribution. For a ball of radius ϵ centered at the origin (where

$1 > \epsilon > 0$), it is easy to check the following inequality:

$$\mu(B_\epsilon(0)) \geq \frac{2\epsilon}{\sqrt{2\pi\sigma}} e^{-\frac{(1-\epsilon)^2}{2\sigma^2}} \geq \frac{2\epsilon}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}}.$$

On the other hand, the Lebesgue measure (λ) in \mathbb{R}^2 assigns volume $\lambda(B_\epsilon(0)) = \pi\epsilon^2$. Clearly, $\frac{d\mu}{d\lambda}$ blows up at the origin. Thus although the center of the circle is “infinitely likely,” it needs to be discarded to recover the homology of the circle. This can only be done by choosing with care the size of the neighborhoods for cleaning the data.

4. As we will elaborate in a later section, the homology finding algorithm can itself be implemented as a spectral algorithm on simplicial complexes. In spectral clustering, one typically constructs a graph from the data and uses the graph Laplacian to partition the graph and thus the data. In our case, since we are interested not just in the number of partitions (clusters) but also the topological nature of these clusters (e.g., circle versus point in the example figures before), we will need to compute the higher homologies. This involves the construction of a suitable simplicial complex from the data. The combinatorial Laplacian on this complex provides the higher homologies. In this sense, our algorithm may be viewed as a generalized form of spectral learning.
5. Here and below, it is possible to simplify considerably the calculation of homology by producing a smaller simplicial complex using a subset of the data. The idea is this: The number of sample points chosen is to guarantee that the manifold is well covered by balls around these points. However, as in the coupon collector problem, one randomly tends to oversample certain regions on the way to coverage. There is no reason to use all of these “extra” sample points in producing the simplicial complex. Choosing a minimal covering set from the data does not lower the sample complexity at all, but it makes the associated simplicial complex simpler, and the boundary maps in the chain complex into reasonably sparse matrices.

Our main theorem exploits the concentration properties of the Gaussian distribution in high dimensional spaces. More generally, one can consider probability distributions that can be factored into a part with a geometric core (on the manifold) and a part in the normal directions (off the manifold) following section 2.1. In this more general setting, one can prove the following theorem.

THEOREM 2. *Let \mathcal{M} be a compact submanifold of \mathbb{R}^N with condition number τ , and let μ be a probability measure on \mathbb{R}^N satisfying the following conditions:*

- (1) *There is an $s > 0$ and a positive real number α so that $\mu(B_s(q)) > \alpha\mu(B_s(p))$ for any $q \in \mathcal{M}$ and any p .*
- (2) *There is a positive real number $\beta < 1$ so that $\mu(B_s(p)) < \alpha\beta\mu(B_s(q))$ if $q \in \mathcal{M}$ and $d(p, \mathcal{M}) > 2s$.*
- (3) *In addition, $s < \tau/5$.*
- (4) *There is a $C > 0$ so that $\mu(B_C(0)) > \frac{1}{2}$.*

Then it is possible to give an algorithm that computes from many μ -random samples the homology (and homotopy type) of \mathcal{M} . The probability of error can be estimated in terms of $(n, \tau, s, C, \alpha, \beta)$.

Note, of course, that the existence of a C in (4) is automatic. However, it is not possible to give a bound on it in terms of the other parameters. Essentially, it is related to the problem of “large diffuse dust clouds”: almost all of the mass of μ can

be concentrated on points of probability almost 0; as a result, it would be very hard to find the much stronger “signal” of \mathcal{M} . Note, too, that for very reasonable measures, condition (2) is unlikely to hold for very small values of s because of the “hotspot” example mentioned above. Part of the proof of the main theorem is checking that in the situation of Gaussian noise, for any s , controlling the variance suffices to ensure (2).

The proof of this is rather easier than the proof of the main theorem and follows the same outline. Essentially, one uses the tension between (1) and (2) to devise a test to eliminate “less likely balls” to clean the data. One estimates, by the techniques of [17] and some elementary proof of the law of large numbers, the probability of including spurious balls (e.g., ones centered outside a $2s$ neighborhood of \mathcal{M}) and that one has covered \mathcal{M} . When one is done, one takes the nerve of covering by balls of size $4s$ that were allowed in, and, by the results of [17], we get a computation of the homotopy type of \mathcal{M} . Finally, it is worth noting that while these arguments allow us to handle a more general setting than our main theorem, unfortunately the complexity of the algorithm for this more general case depends exponentially on N , the ambient dimension.

Remark. It is worth being explicit about the meaning of the statement of what it means to compute a homotopy type: all that is asserted is that one has produced a simplicial complex in the correct homotopy type. From this, it is easy to compute the fundamental group (as a group presentation: although this determines the group up to isomorphism, it is not possible to determine algorithmically whether or not it is trivial) and the homology groups. Higher homotopy is determined by such a description, but there is no straightforward method of computation (even in the weak form of generators and relations). If one has additional information, e.g., that $\dim \mathcal{M} = 2$, then the homology information determines the diffeomorphism type of the surface, but as the dimension increases, the difficulty of inferring useful information from such a description increases quickly.

3. Basic algorithmic framework. The basic algorithmic framework consists of drawing a number of points according to P , filtering these points according to a “cleaning procedure,” and then constructing a simplicial complex according to the constructions outlined in [17]. The detailed steps of this framework are as follows:

1. Draw a set of n points $\bar{x} = \{x_1, \dots, x_n\}$ in i.i.d. fashion from P .
2. Construct the nearest neighbor graph with n vertices (each vertex associated to a data point) with adjacency matrix

$$W_{ij} = 1 \iff \|x_i - x_j\| < s.$$

Thus two points x_i and x_j are connected if $B_{s/2}(x_i)$ and $B_{s/2}(x_j)$ intersect.

3. Let d_i be the degree of the i th vertex (associated to x_i) of the graph. Throw away all data points whose degree is smaller than some prespecified threshold. Let the filtered set be $\bar{z} \subset \bar{x}$.
4. With the points that are left, construct the set

$$U = \cup_{x \in \bar{z}} B_\epsilon(x).$$

5. Compute the homology of U by constructing the simplicial complex K corresponding to the nerve of U according to [17].

In the above framework, there is a one-step cleaning procedure. Many variants of the above framework may be considered. Conceptually, if we are able to filter

out points in a neighborhood of the medial axis, retaining sufficient points in the neighborhood of the manifold, then we should be able to reconstruct the manifold suitably. In this paper we provide some details as to how this might be done for reconstructing the homology of the manifold.

3.1. Remarks and elaborations. In step 2 of the above algorithm, a choice has to be made for s . As we shall see, a suitable choice is $s = 4r$, where r is a number satisfying

$$\sqrt{8(D-d)}\sigma < r \quad \text{and} \quad 9r < (\sqrt{9} - \sqrt{8})\tau.$$

Such a number always exists by assumption on the noise in 1. In step 3 of the algorithm, a choice has to be made for the value of the threshold to prune the set of data. A suitable choice here is

$$d_i > \frac{3a}{4}(1 - (r/2\tau)^2)^{d/2} \text{vol}(B_r^d).$$

In step 5 of the algorithm, one constructs the nerve of U at a scale ϵ . This ϵ is different from s chosen earlier (a value $\epsilon = \frac{9r+\tau}{2}$ suffices, but details will be specified in subsequent developments and in the propositions stated later). The nerve of U is a simplicial complex constructed as follows. The j -skeleton consists of j -simplices where each j -simplex consists of $j+1$ distinct points $z_1, \dots, z_{j+1} \in \bar{z}$ such that $\cap B_\epsilon(z_i) \neq \phi$. The 1-skeleton of this complex is therefore a graph where the vertices are identified with the data (after cleaning) and the edges link two points which are 2ϵ -close.

3.1.1. The combinatorial Laplacian and its kernel. The homology of the manifold is obtained by computing the Betti numbers of the data-derived simplicial complex. This, in turn, is obtained from the eigenspaces of the combinatorial Laplacian defined on this complex (see [13]). Thus, there is a natural spectral algorithm for our problem that is a generalization of spectral clustering used in determining the number of clusters of a data set. Let us elaborate.

1. One begins by picking an orientation for the complex as follows. Recall that every j -simplex $\sigma \in K$ is associated with a set of $j+1$ points. If $x_{i_0}, x_{i_1}, \dots, x_{i_j}$ (where i_l 's are in increasing order) are the set of points associated with a particular j -simplex μ , then an orientation for this is picked by choosing an ordering of the vertices. A possible choice is to simply choose the ordering given by $[i_0 i_1 \dots i_j]$. Therefore if $\mu = [i_0 i_1 \dots i_j]$ and π is a permutation of $0, \dots, j$, then $\text{sign}(\pi)\mu = [i_{\pi(0)} i_{\pi(1)} \dots i_{\pi(j)}]$.

As an example, every 1-simplex is an edge and one picks an orientation by picking an ordering of the vertices that define the edge.

2. A j -chain is defined as a formal sum

$$c = \sum_{\mu} \alpha_{\mu} \mu,$$

where $\alpha_{\mu} \in \mathbb{R}$ and $\mu \in K_j$ (the set of j -simplices) is a j -simplex. Let C_j be the vector space of j -chains. This is a vector space of dimensionality $n_j = |K_j|$.

3. The boundary operators are defined as

$$\partial_j : C_j \rightarrow C_{j-1}$$

in the following way. For any $\mu \in K_j$ (corresponding to the oriented simplex $[x_{i_0} \dots x_{i_j}]$), we have $\partial_j(\mu) = \sum_{i=0}^j (-1)^i \mu_i$, where μ_i is the oriented $j-1$ -simplex corresponding to $[x_{i_0} x_{i_1} \dots x_{i_{i-1}} x_{i_{i+1}} \dots x_{i_j}]$. By linearity, for any

$c = \sum_{\mu} \alpha_{\mu} \mu$, we have

$$\partial_j(c) = \sum_{\mu} \alpha_{\mu} \partial_j(\mu).$$

The boundary operators are therefore linear operators that can be represented as $n_{j-1} \times n_j$ matrices. The adjoint is therefore defined as

$$\partial_j^* : C_{j-1} \rightarrow C_j$$

and can be represented as an $n_j \times n_{j-1}$ matrix.

4. The combinatorial Laplacian Δ_j for each j is

$$\Delta_j = \partial_j^* \partial_j + \partial_{j+1} \partial_{j+1}^*.$$

Clearly

$$\Delta_j : C_j \rightarrow C_j.$$

5. The Betti numbers b_0, b_1, \dots are obtained as the null space of Δ_j .

Remark. It is worth noting that with the definitions above,

$$\Delta_0 : C_0 \rightarrow C_0$$

corresponds to the standard graph Laplacian where C_0 is the set of functions on the vertex set of the complex and C_1 is the set of edges (1-simplices) of the complex. This kind of a nearest neighbor graph is often constructed in spectral applications. Indeed, the dimensionality of the nullspace of Δ_0 gives the number of connected components of the graph (b_0) which in turn is related to the number of connected components of the underlying manifold. The number of connected components is usually interpreted as the number of clusters in the spectral clustering tradition (see, for example, Ding and Zha [10] and the references therein).

4. Analysis of the cleaning procedure. We prove the following $A - B$ lemma that is important in analyses of the cleaning procedure generally. Imagine we have two sets $A, B \subset \mathbb{R}^D$ such that

$$\inf_{x \in A, y \in B} \|x - y\| = \eta > 0.$$

The larger η is, the easier it should be to separate samples of A from samples from B . Let there be a probability measure μ on \mathbb{R}^D according to which points $x_1, \dots, x_n \in \mathbb{R}^D$ are drawn. We define (for $s > 0$, in particular $s \ll \eta/2$)

$$\alpha_s = \inf_{p \in A} \mu(B_s(p))$$

and

$$\beta_s = \sup_{p \in B} \mu(B_s(p)).$$

Then the following is true.

LEMMA 3. *Let $\alpha_s \geq \alpha > \beta \geq \beta_s$ and set $h = \frac{\alpha - \beta}{2}$. Then if the number of points n is greater than $4\eta \log(\eta)$, where*

$$\eta = \max \left(\left(1 + \frac{2}{h^2} \log \left(\frac{2}{\delta} \right) \right), 4 \right),$$

then, with probability $> 1 - \delta$, both (i) and (ii) below are true:

- (i) for all $x_i \in A$, $\frac{d_i}{n-1} > \frac{\alpha+\beta}{2}$;
(ii) for all $x_j \in B$, $\frac{d_j}{n-1} < \frac{\alpha+\beta}{2}$,

where $d_i = \sum_{j \neq i} 1_{[x_j \in B_s(x_i)]}$.

Proof. Given the random variables x_1, \dots, x_n , we define an event A_i as follows. Consider the random variables $y_{i,j} = 1_{[x_j \in B_s(x_i)]}$. Then for each $j \neq i$, the $y_{i,j}$'s are 0–1-valued and i.i.d. with mean $\mu(B_s(x_i))$. Define the event A_i to be

$$A_i : \left| \frac{1}{n-1} \sum_{j \neq i} y_{i,j} - \mu(B_s(x_i)) \right| > h.$$

By a simple application of Chernoff's inequality, we have

$$\mathbb{P}(A_i) < 2e^{-\frac{h^2(n-1)}{2}}.$$

By the union bound,

$$\mathbb{P}(\cup_i A_i) < \sum_{i=1}^n \mathbb{P}(A_i) = 2ne^{-\frac{h^2(n-1)}{2}}.$$

Therefore, if

$$2ne^{-\frac{h^2(n-1)}{2}} < \delta$$

with probability greater than $1 - \delta$, for all i simultaneously,

$$\left| \frac{d_i}{n-1} - \mu(B_s(x_i)) \right| \leq h.$$

Therefore, if $x_i \in A$, we have

$$\frac{d_i}{n-1} \geq \mu(B_s(x_i)) - h \geq \alpha - h,$$

and if $x_i \in B$, we have

$$\frac{d_i}{n-1} \leq \mu(B_s(x_i)) + h \leq \beta + h.$$

Putting in the value of h , the result follows. Now it only remains to check the bound on n . We see that as long as

$$n-1 > \frac{2}{h^2} (\log(n) + \log(2/\delta)),$$

it follows that

$$n > \left(1 + \frac{2}{h^2} \log(2/\delta)\right) + \frac{2}{h^2} \log(n).$$

For $\delta < \frac{1}{2}$, we have that $(1 + \frac{2}{h^2} \log(2/\delta)) > \frac{2}{h^2}$, so that it is enough to find an n satisfying

$$n > x + x \log(n),$$

where $x = (1 + \frac{2}{h^2} \log(2/\delta))$. The bound for n now follows from a straightforward calculation. \square

5. Analysis of Gaussian noise off the manifold. We apply the “ $A - B$ ” lemma to the case of manifold (condition number τ) and the probability measure μ obtained by pushing down P as described earlier. We choose a number r that satisfies

$$\sqrt{8(D-d)}\sigma < r \quad \text{and} \quad 9r < (\sqrt{9} - \sqrt{8})\tau.$$

Our data are to be cleaned by considering balls of radius $s = 4r$ centered at each of the points, building the associated graph and throwing away points associated to low degree vertices. In order to proceed, we choose (where $R = 9r$)

$$A = \text{Tub}_r(\mathcal{M}) \quad \text{and} \quad B = \mathbb{R}^D \setminus \text{Tub}_R(\mathcal{M}).$$

With these choices, we now lowerbound α_s and upperbound β_s . We will need the following standard concentration lemma for Gaussian probability distributions, which can be proved by the introduction of spherical coordinates.

LEMMA 4. *Let X be a normally distributed random variable with mean 0 and identity covariance matrix I in k -dimensions. If γ_k is the measure (on \mathbb{R}) associated with X , then*

$$\gamma_k \left\{ x \in \mathbb{R}^k \text{ such that } \|x\|^2 \geq k + \delta \right\} = \mathbb{P} \left(\|X\|^2 \geq k + \delta \right) \leq \left(\frac{k}{k + \delta} \right)^{-k/2} e^{-\delta/2}.$$

For our purposes, a more convenient form can be derived as follows.

LEMMA 5. *Let X be a normally distributed random variable with mean 0 and covariance matrix $\sigma^2 I$ in k -dimensions. If ν_k is the measure (on \mathbb{R}^k) associated with X , then for any $T > \sigma^2 k$, we have*

$$\nu_k \left\{ x \in \mathbb{R}^k \text{ such that } \|x\|^2 \geq T \right\} \leq (eze^{-z})^{k/2},$$

where $z = \frac{T}{\sigma^2 k}$.

Proof. Define the random variable $Y = X/\sigma$. Clearly, Y is normally distributed with mean 0 and covariance matrix I . Therefore

$$\gamma_k \left\{ y \in \mathbb{R}^k \text{ such that } \|y\|^2 \geq k + \delta \right\} = \nu_k \left\{ x \in \mathbb{R}^k \text{ such that } \|x\|^2 \geq \sigma^2(k + \delta) \right\}.$$

Put $T = \sigma^2 k + \sigma^2 \delta$. We then have

$$\begin{aligned} \nu_k \left\{ x \in \mathbb{R}^k \text{ such that } \|x\|^2 \geq \sigma^2(k + \delta) \right\} &\leq \left(\frac{k}{k + \delta} \right)^{-k/2} e^{-\delta/2} \\ &= \left(\frac{\sigma^2 k}{T} \right)^{-k/2} e^{-1/2(\frac{T}{\sigma^2} - k)} = \left(\frac{eT}{\sigma^2 k} e^{-T/(\sigma^2 k)} \right)^{k/2}. \quad \square \end{aligned}$$

5.1. Lower bounding α_s . Consider some $x \in A$ and $\mu(B_s(x))$ for such an x . Let $p = \pi_0(x) \in \mathcal{M}$. Then since $\|x - p\| < r$, we clearly have

$$\mu(B_s(x)) \geq \mu(B_{2r}(p)).$$

In turn, we have

$$\mu(B_{2r}(p)) \geq \int_{x \in \mathcal{M} \cap B_r(p)} P(x) \int_{B_r^{D-d} \in T_x^\perp} P(y|x).$$

For any $x \in \mathcal{M}$, the probability distribution $P(y|x)$ is normally distributed and we are in a position to apply a concentration equality for Gaussians. We see that

$$\int_{B_r^{D-d} \in T_x^\perp} P(y|x) = 1 - \gamma,$$

where

$$\gamma = \mathbb{P}(\|y\| > r^2) \leq \left(\frac{er^2}{(D-d)\sigma^2} \right)^{(D-d)/2} e^{-\frac{r^2}{2\sigma^2}}.$$

Therefore, we have

$$\mu(B_{2r}(p)) \geq (1 - \gamma) \int_{x \in \mathcal{M} \cap B_r(p)} P(x) \geq a(1 - \gamma) \text{vol}(B_r(p) \cap \mathcal{M}).$$

Since the curvature of \mathcal{M} is bounded by the τ constraint (see Lemma 5.3 on p. 430 of [17]), we have that

$$\text{vol}(B_r(p) \cap \mathcal{M}) \geq (1 - (r/2\tau)^2)^{d/2} \text{vol}(B_r^d),$$

where $\theta = \arcsin\left(\frac{r}{2\tau}\right)$.

Thus we have

$$\alpha_s > a(1 - \gamma)(1 - (r/2\tau)^2)^{d/2} \text{vol}(B_r^d).$$

It is worth noting that as $D - d$ (the codimension) gets large, or as σ gets small, the quantity γ decreases eventually to zero.

5.2. Upperbounding β_s . Consider some $z \in B = \mathbb{R}^D \setminus \text{Tub}_R(\mathcal{M})$. Noting that $s = 4r$ and $R = 9r$, we see

$$\mu(B_s(z)) \leq \mu(\mathbb{R}^D \setminus \text{Tub}_{R-s}(\mathcal{M})) = \mu(\mathbb{R}^D \setminus \text{Tub}_{5r}(\mathcal{M})).$$

Now

$$\mu(\mathbb{R}^D \setminus \text{Tub}_{5r}(\mathcal{M})) = \int_{x \in \mathcal{M}} P(x) \int_{y \notin B_{5r} \cap T_x^\perp} P(y|x).$$

By the concentration of inequality of Gaussians, we have

$$\int_{y \notin B_{5r} \cap T_x^\perp} P(y|x) \leq \left(\frac{25er^2}{(D-d)\sigma^2} \right)^{\frac{D-d}{2}} e^{-\frac{25r^2}{2\sigma^2}}.$$

Therefore, it follows that

$$\beta_s \leq \left(\frac{25er^2}{(D-d)\sigma^2} \right)^{\frac{D-d}{2}} e^{-\frac{25r^2}{2\sigma^2}}.$$

5.3. A density lemma. In this section, we provide a bound on how many points we need to draw in order to obtain a suitably dense set of points in a neighborhood of \mathcal{M} . Begin by letting p_1, p_2, \dots, p_l be a set of points in \mathcal{M} such that $M \subset \cup_{i=1}^l B_r(p_i)$. In [17], an upper bound was derived on the size of the cover l as

$$l \leq \frac{\text{vol}(\mathcal{M})}{\cos^d(\theta)\text{vol}(B_r^d)},$$

where $\theta = \arcsin(\frac{r}{2r})$.
 Now we note that

$$\mu(B_r(p_i)) \geq \int_{x \in \mathcal{M} \cap B_{r/2}(p_i)} P(x) \int_{B_{r/2}^{D-d} \cap T_x^\perp} P(y|x).$$

For each x , following the usual arguments, we have

$$\int_{B_{r/2}^{D-d} \cap T_x^\perp} P(y|x) = 1 - \gamma,$$

where by the Gaussian concentration inequality, we have

$$\gamma = \mathbb{P}\left(\|y\|^2 > \frac{r^2}{4}\right) \leq \left(\frac{er^2}{4(D-d)\sigma^2}\right)^{(D-d)/2} e^{-\frac{r^2}{8\sigma^2}}.$$

We can now state the following lemma, which guarantees that if we draw enough points, we will get a suitably dense set of points in a neighborhood of \mathcal{M} .

LEMMA 6. *Let $A_i = B_r(p_i)$ such that $\cup_{i=1}^l A_i$ form a suitable minimal cover of \mathcal{M} . Let $\mu(A_i) \geq t$. Let $\bar{x} = \{x_1, \dots, x_n\}$ be a collection of n data points drawn in i.i.d. fashion according to μ .*

Then, if $n > \frac{1}{t}(\log(l) + \log(2/\delta))$, with probability greater than $1 - \delta/2$, each A_i has at least one data point in it, i.e., for all i , $A_i \cap \bar{x} \neq \phi$.

5.4. Putting it all together. We begin by using the following simple lemma.

LEMMA 7. *The function $f(x) = (xe^{-x})^n$ is a decreasing function of x for $x > 1$. Thus, we have that*

$$\forall x \geq 2, f(x) < \left(\frac{2}{e}\right)^n < \sqrt{\frac{2}{e}}.$$

Further, for all $x > x_* > 1$,

$$f(x) \leq \left(\frac{x_*}{e^{x_*-1}}\right)^n.$$

Proof. This is simply proved by noting that $f'(x) < 0$ for all $x > 1$. □

We now see that if the codimension $D - d$ is sufficiently high, then $\beta < \alpha/2$. This is stated as follows.

LEMMA 8. *Letting*

$$D - d \geq \frac{1}{196} \left(\log\left(\frac{4}{a}\right) + d \log\left(\frac{1}{\cos(\theta)}\right) + \log\left(\frac{1}{\text{vol}(B_r^d)}\right) \right),$$

we have that $\beta_s \leq \beta < \alpha/2 \leq \frac{\alpha_}{2}$.*

Proof. First consider α_s . By the previous argument in section 5.1, we see that

$$\alpha_s > \alpha = a(1 - \gamma) \cos^d(\theta) \text{vol}(B_r^d),$$

where $\gamma < (eze^{-z})^{\frac{D-d}{2}}$ with $z = \frac{r^2}{(D-d)\sigma^2}$. Since by our main assumption we have that $r > \sqrt{8}\sigma\sqrt{D-d}$, we see that $z > 8$. By Lemma 7, we have that $\gamma < \sqrt{2/e}$ and so $(1 - \gamma) > 1 - \sqrt{2/e}$. In fact, since $z > 8$ in this case, it turns out that $\gamma < \frac{1}{2}$ so that

$$\alpha > \frac{a}{2} \cos^d(\theta) \text{vol}(B_r^d).$$

Now consider β_s . By the argument in section 5.2, we have that

$$\beta_s \leq \beta = (eze^{-z})^{\frac{D-d}{2}},$$

where $z = \frac{25r^2}{(D-d)\sigma^2}$. Again, by our main assumption, we see that $z \geq 400$ so that

$$\beta \leq (e \cdot 400 \cdot e^{-400})^{(D-d)/2}.$$

Therefore, for β to be less than $\alpha/2$, it is sufficient for

$$(e \cdot 400 \cdot e^{-400})^{(D-d)/2} \leq \frac{a}{4} \cos^d(\theta) \text{vol}(B_r^d).$$

Taking logs, we get that it is sufficient for

$$(D - d)/2 \log(e^{399}/400) \geq \log\left(\frac{4}{a}\right) + d \log\left(\frac{1}{\cos(\theta)}\right) + \log\left(\frac{1}{\text{vol}(B_r^d)}\right).$$

Noting that $e^{399}/400 > e^{392}$, we see it is sufficient for

$$D - d > \frac{1}{196} \left(\log\left(\frac{4}{a}\right) + d \log\left(\frac{1}{\cos(\theta)}\right) + \log\left(\frac{1}{\text{vol}(B_r^d)}\right) \right).$$

The result follows. \square

We are now in a position to prove a central proposition.

PROPOSITION 9. *Let μ be a probability measure on \mathbb{R}^D satisfying the strong variance condition described in section 1. Let $\bar{x} = \{x_1, \dots, x_n\}$ be n data points drawn in i.i.d. fashion according to μ . Let $\bar{x}' \subset \bar{x}$ be the data points left after cleaning by the procedure described previously.*

Then if $n > \max(A, B)$, with probability greater than $1 - \delta$, \bar{x}' is a set of points that (i) is in the R -tubular neighborhood of \mathcal{M} , i.e., $\bar{x}' \subset \text{Tub}_R(\mathcal{M})$, and (ii) is R -dense in \mathcal{M} , i.e., $\mathcal{M} \subset \cup_{x \in \bar{x}'} B_R(x)$. Here

$$A = 4y \log(y), \text{ where } y = \max(4, x); x = 1 + \frac{32}{a^2(1 - (r/2\tau)^2)^d \text{vol}^2(B_r^d)} \log\left(\frac{4}{\delta}\right),$$

and

$$B = \frac{1}{(1 - \sqrt{\frac{2}{e}})(1 - (r/4\tau)^2)^{d/2} \text{vol}(B_{r/2}^d)} \left(\log\left(\frac{2\text{vol}(\mathcal{M})}{\delta(1 - (r/2\tau)^2)^{d/2} \text{vol}(B_r^d)}\right) \right).$$

Proof. We show (in Part 1 below) that if $n > A$, with probability greater than $1 - \delta/2$, \bar{x}' is such that (i) all the points in \bar{x} that are in $\text{Tub}_r(\mathcal{M})$ are retained in \bar{x}' , and (ii) no points of \bar{x} that are in $\text{Ext}(\text{Tub}_R(\mathcal{M}))$ are retained in \bar{x}' . We will prove this by an application of the “ $A - B$ ” lemma and a calculation of the precise bounds.

We then show (Part 2) that if $n > B$, with probability greater than $1 - \delta/2$, \bar{x} is such that $\mathcal{M} \subset \cup_{x \in \bar{x} \cup \text{Tub}_r(\mathcal{M})} B_{2r}(x)$. We will prove this by an application of the density lemma and a calculation of the precise bounds. Taken together, Parts 1 and 2 show that if $n > \max(A, B)$, with probability greater than $1 - \delta$, the following facts are true: (i) $\bar{x}' \subset \text{Tub}_R(\mathcal{M})$, (ii) $\mathcal{M} \subset \cup_{x \in \bar{x}'} B_{2r}(x) \subset \cup_{x \in \bar{x}'} B_R(x)$. The proposition follows immediately.

Part 1. By Lemma 8, we see that $\beta < \alpha/2$. Therefore, $h = \alpha - \beta > \alpha/2$. By applying the “ $A - B$ ” lemma, and choosing $A = \text{Tub}_r(\mathcal{M})$ and $B = \text{Ext}(\text{Tub}_R(\mathcal{M}))$, we see that if $n > 4y \log(y)$, where $y = \max(4, x)$ and

$$x = \left(1 + \frac{8}{\alpha^2} \log \left(\frac{4}{\delta} \right) \right),$$

the cleaning procedure will retain all points in $\bar{x} \cap \text{Tub}_r(\mathcal{M})$ while eliminating all points in $\bar{x} \cap \text{Ext}(\text{Tub}_R(\mathcal{M}))$. By the calculations in section 5.1 and the proof of Lemma 8, we see that

$$\alpha > \frac{a}{2} (1 - (r/2\tau)^2)^{d/2} \text{vol}(B_r^d).$$

Putting this in proves Part 1.

Part 2. Here we apply Lemma 6. Let us bound t and l of that lemma appropriately. Note that by the arguments presented in section 5.1, we have

$$\mu(A_i) \geq a(1 - \gamma)(1 - (r/4\tau)^2)^{d/2} \text{vol}(B_{r/2}^d),$$

where $\gamma \leq (eze^{-z})^{(D-d)/2}$ such that $z = \frac{r^2}{4(D-d)\sigma^2} \geq 2$. Then by Lemma 7 we have that

$$\mu(A_i) \geq \left(1 - \sqrt{\frac{2}{e}} \right) (1 - (r/4\tau)^2)^{d/2} \text{vol}(B_{r/2}^d).$$

Similarly, for l , we have that $l \leq \frac{\text{Vol}(\mathcal{M})}{(1 - (r/4\tau)^2)^{d/2} \text{vol}(B_{r/2}^d)}$. Therefore

$$\log(l) \leq \log(\text{vol}(\mathcal{M})) + d \log \left(\frac{1}{(1 - (r/2\tau)^2)^{1/2}} \right) + \log \left(\frac{1}{\text{vol}(B_r^d)} \right).$$

Putting these together proves Part 2. □

Thus we see that the procedure yields a set of points \bar{x} that are R -dense in \mathcal{M} and entirely contained in an R -tube around \mathcal{M} . At this point, we can invoke the following proposition proved in [17].

PROPOSITION 10. *Let S be a set of points in the tubular neighborhood of radius R around \mathcal{M} . Let U be given by*

$$U = \cup_{x \in S} B_\epsilon(x).$$

If S is R -dense in \mathcal{M} , then \mathcal{M} is a deformation retract of U for all $R < (\sqrt{9} - \sqrt{8})\tau$ and $\epsilon \in \left(\frac{(R+\tau) - \sqrt{R^2 + \tau^2 - 6\tau R}}{2}, \frac{(R+\tau) + \sqrt{R^2 + \tau^2 - 6\tau R}}{2} \right)$.

Now we can prove our main theorem by combining Propositions 9 and 10.

THEOREM 11. *Let $\bar{x} = \{x_1, \dots, x_n\}$ be a collection of n i.i.d. points drawn at random from μ satisfying the strong variance condition described earlier. Then, as long as $n > \max(A, B)$ with probability greater than $1 - \delta$, the algorithm described earlier returns the homology of \mathcal{M} .*

6. Conclusions. In this paper, we have taken a topological view of unsupervised learning from data in a high dimensional space. Our approach is conditioned on the belief that in high dimensions, “natural” probability distributions are far from uniform and concentrate around lower dimensional structures. To model this intuition we consider a probability distribution that concentrates around a low dimensional submanifold of the ambient space with noise in the normal directions. With random data drawn from such a probability distribution, we show that if the noise is sufficiently small, it is possible to recover the homology of the manifold by a spectral algorithm that depends very weakly on the ambient dimension. This result is a step towards a larger understanding of when we might expect to make effective inferences in high dimensional spaces without running into the curse of dimensionality.

REFERENCES

- [1] N. AMENTA AND M. BERN, *Surface reconstruction by Voronoi filtering*, Discrete Comput. Geom., 22 (1999), pp. 481–504.
- [2] N. AMENTA, S. CHOI, T. DEY, AND N. LEEKHA, *A simple algorithm for homeomorphic surface reconstruction*, in Proceedings of the Sixteenth Annual Symposium on Computational Geometry, ACM, New York, 2000, pp. 213–222.
- [3] M. BELKIN AND P. NIYOGI, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396.
- [4] M. BELKIN AND P. NIYOGI, *Towards a theoretical foundation for Laplacian-based manifold methods*, J. Comput. System Sci., 74 (2008), pp. 1289–1308.
- [5] J. BOISSONNAT, L. GUIBAS, AND S. OUDOT, *Manifold reconstruction in arbitrary dimensions using witness complexes*, Discrete Comput. Geom., 42 (2009), pp. 37–70.
- [6] F. CHAZAL AND A. LIEUTIER, *Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees*, Comput. Geom., 40 (2008), pp. 156–170.
- [7] S.-W. CHENG, T. K. DEY, AND E. A. RAMOS, *Manifold reconstruction from point samples*, in Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2005, pp. 1018–1027.
- [8] D. COHEN-STEINER, H. EDELSBRUNNER, AND J. HARER, *Stability of persistence diagrams*, Discrete Comput. Geom., 37 (2007), pp. 103–120.
- [9] S. DASGUPTA, *Learning mixtures of Gaussians*, in Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society Press, 1999, pp. 634–644.
- [10] C. DING AND H. ZHA, *Spectral Clustering, Ordering and Ranking: Statistical Learning with Matrix Factorizations*, Springer, New York, 2011.
- [11] H. EDELSBRUNNER AND J. HARER, *Persistent homology—a survey*, in Surveys on Discrete and Computational Geometry: Twenty Years Later: AMS-IMS-SIAM Joint Summer Research Conference (Snowbird, UT, 2006), Vol. 453, AMS, Providence, RI, 2008, pp. 257–282.
- [12] H. EDELSBRUNNER AND E. MÜCKE, *Three-dimensional alpha shapes*, ACM Trans. Graphics, 13 (1994), pp. 43–72.
- [13] J. FRIEDMAN, *Computing Betti numbers via combinatorial Laplacians*, Algorithmica, 21 (1998), pp. 331–346.
- [14] E. GINÉ AND V. KOLTCHINSKII, *Empirical graph Laplacian approximation of Laplace-Beltrami operators: Large sample results*, in High Dimensional Probability, IMS Lecture Notes Monogr. Ser. 51, Inst. Math. Statist., Beachwood, OH, 2006, pp. 238–259.
- [15] M. HEIN, J. AUDIBERT, AND U. VON LUXBURG, *From graphs to manifolds: Weak and strong pointwise consistency of graph Laplacians*, in Learning Theory, Lecture Notes in Comput. Sci. 3559, Springer, Berlin, 2005, pp. 470–485.
- [16] T. KACZYŃSKI, K. MISCHAIKOW, AND M. MROZEK, *Computational Homology*, Springer, New York, 2004.

- [17] P. NIYOGI, S. SMALE, AND S. WEINBERGER, *Finding the homology of submanifolds with high confidence from random samples*, Discrete Comput. Geom., 39 (2008), pp. 419–441.
- [18] A. SINGER, *From graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134.
- [19] S. VEMPALA AND G. WANG, *A spectral algorithm for learning mixtures of distributions*, J. Comput. System Sci., (2002), pp. 113–123.
- [20] A. ZOMORODIAN AND G. CARLSSON, *Computing persistent homology*, Discrete Comput. Geom., 33 (2005), pp. 249–274.