Stratification Learning through Homology Inference

Paul Bendich
Institute of Science and Technology Austria
Klosterneuburg, Austria
bendich@ist.ac.at

Sayan Mukherjee
Department of Statistical Science
Duke University, Durham NC, USA
sayan@stat.duke.edu

Bei Wang
Department of Computer Science
Duke University, Durham NC, USA
beiwang@cs.duke.edu

Abstract

We develop a topological approach to stratification learning. Given point cloud data drawn from a stratified space, our objective is to infer which points belong to the same strata. First we define a multi-scale notion of a stratified space, giving a stratification for each radius level. We then use methods derived from kernel and cokernel persistent homology to cluster the data points into different strata, and we prove a result which guarantees the correctness of our clustering, given certain topological conditions. We later give bounds on the minimum number of sample points required to infer, with probability, which points belong to the same strata. Finally, we give an explicit algorithm for the clustering and apply it to some simulated data.

1 Introduction

Manifold learning is a basic problem in geometry, topology, and statistical inference that has received a great deal of attention. The basic idea is as follows: given a point cloud of data sampled from a manifold in an ambient space $\mathbb{R}^N$, infer the underlying manifold. A limitation of the problem statement is that it does not apply to sets that are not manifolds. For example, we may consider the more general class of stratified spaces that can be decomposed into strata, which are manifolds of varying dimension, each of which fit together in some uniform way inside the higher dimensional space.

In this paper, we study the following problem in stratification learning: given a point cloud sampled from a stratified space, how do we cluster the points so that points in the same cluster are in the same stratum, while points in different clusters are not? Intuitively, the strategy should be clear: two points belong in the same stratum if they "look the same locally," meaning that they have identical neighborhoods, within the larger space, at some very small scale. However, the notion of "local" becomes unclear in the context of sampling uncertainty, since everything becomes quite noisy at vanishingly small scale. In response, we introduce a radius parameter r and define a notion of local equivalence at each such r.

Our tools are derived from algebraic topology. In particular, we define local equivalence between points via maps between relative homology groups, and we then attempt to infer this relation by using ideas coming from persistent homology (Edelsbrunner and Harer 2010).

Prior Work The field of topological data analysis is expanding. Topological persistence has been used to analyze scaler fields over point cloud data (Chazal et al. 2009b) and methods have been developed that reduce high dimensional data sets into simplicial complexes that capture the topological and geometric information (Singh, Mémoli, and Carlsson 2007).

Consistency in manifold learning has often been recast as a homology inference statement: as the number of points in a point cloud goes to infinity, the inferred homology converges to the true homology of the underlying space. Results of this nature have been given for manifolds (Niyogi, Smale, and Weinberger 2008a; 2008b) and a large class of compact subsets of Euclidean space (Chazal, Cohen-Steiner, and Lieutier 2009). Stronger results in homology inference for closed subsets of a metric space are given in (Cohen-Steiner, Edelsbrunner, and Harer 2007).

Geometric approaches to stratification inference have been developed including inference of a mixture of linear subspaces (Lerman and Zhang 2010), mixture models for general stratified spaces (Haro, Randall, and Sapiro 2007), and generalized Principal Component Analysis (GPCA) (Vidal, Ma, and Sastry 2005) developed for dimension reduction for mixtures of manifolds.

The study of stratified spaces has long been a focus of pure mathematics (Goresky and MacPherson 1988; Weinberger 1994). A deterministic analysis of inference of local homology groups of a stratified space was addressed in (Bendich et al. 2007). An extended version of this paper can be found in (Bendich, Mukherjee, and Wang 2010).

2 Background

We review necessary background on persistent homology and stratified spaces.

Persistence Modules

Let A be some subset of $\mathbb{R}$. A *persistence module* $\mathcal{F}_A$ is a collection $\{F_\alpha\}_{\alpha \in A}$ of $\mathbb{Z}/2\mathbb{Z}$-vector spaces, together with

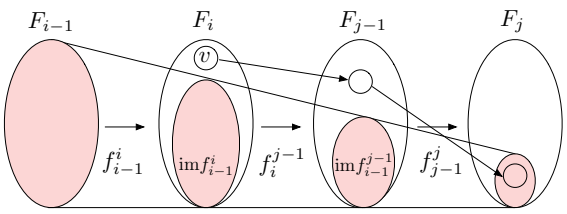

Figure 1: The vector v is born at level i and then it dies at level j.

a family $\{f^\beta_\alpha : F_\alpha \to F_\beta\}_{\alpha \le \beta \in A}$ of linear maps such that $\alpha \le \beta \le \gamma$ implies $f^\gamma_\alpha = f^\gamma_\beta \circ f^\beta_\alpha$ (Chazal et al. 2009a). We will assume that the index set A is either $\mathbb{R}$ or $\mathbb{R}_{\ge 0}$ and not explicitly state indices unless necessary.

A real number α is said to be a *regular value* of the persistence module $\mathcal{F}$ if there exists some $\epsilon > 0$ such that the map $f^{\alpha+\delta}_{\alpha-\delta}$ is an isomorphisms for each $\delta < \epsilon$. Otherwise we say that α is a *critical value* of the persistence module; if $A = \mathbb{R}_{\ge 0}$, then $\alpha = 0$ will always be considered to be a critical value. We say that $\mathcal{F}$ is *tame* if it has a finite number of critical values and if all the vector spaces F_α are of finite rank. Any tame $\mathbb{R}_{\ge 0}$-module $\mathcal{F}$ must have a smallest non-zero critical value $\rho(\mathcal{F})$; we call this number the *feature size* of the persistence module.

Assume $\mathcal{F}$ is tame and so we have a finite ordered list of critical values $0 = c_0 < c_1 < \ldots < c_m$. We choose regular values $\{a_i\}_{i=0}^m$ such that $c_{i-1} < a_{i-1} < c_i < a_i$ for all $1 \le i \le m$, and we adopt the shorthand notation $F_i \equiv F_{a_i}$ and $f^j_i : F_i \to F_j$, for $0 \le i \le j \le m$. A vector $v \in \mathcal{F}_i$ is said to be *born* at level i if $v \notin \mathrm{im}\, f^i_{i-1}$, and such a vector *dies* at level j if $f^j_i(v) \in \mathrm{im}\, f^j_{i-1}$ but $f^{j-1}_i(v) \notin \mathrm{im}\, f^{j-1}_{i-1}$. This is illustrated in Figure 1. We then define $P^{i,j}$ to be the vector space of vectors that are born at level i and then subsequently die at level j, and $\beta^{i,j}$ denotes its rank.

Persistence Diagrams The information contained within a tame module $\mathcal{F}$ is often compactly represented by a *persistence diagram*, $\mathrm{Dgm}(\mathcal{F})$. This diagram is a multi-set of points in the extended plane. It contains $\beta^{i,j}$ copies of the points (c_i, c_j), as well as infinitely many copies of each point along the major diagonal $y = x$. In Figure 2 the persistence diagrams for a curve and a point cloud sampled from it are displayed; see Section 2 for a full explanation of this figure.

For any two points $u = (x, y)$ and $u' = (x', y')$ in the extended plane, we define $||u - u'||_\infty = \max\{|x - x'|, |y - y'|\}$. We define the *bottleneck distance* between any two persistence diagrams D and D' to be:

$$d_B(D, D') = \inf_{\Gamma : D \to D'} \sup_{u \in D} ||u - \Gamma(u)||_\infty,$$

where Γ ranges over all bijections from D to D'. Under certain conditions, persistence diagrams will be stable under the bottleneck distance.

(Co)Kernel Modules Suppose now that we have two persistence modules $\mathcal{F}$ and $\mathcal{G}$ along with a family of maps $\{\phi_\alpha : F_\alpha \to G_\alpha\}$ which commute with the module maps – for every pair $\alpha \le \beta$, we have $g^\beta_\alpha \circ \phi_\alpha = \phi_\beta \circ f^\beta_\alpha$. Then, for each pair of real numbers $\alpha \le \beta$, the restriction of f^β_α to $\ker \phi_\alpha$ maps into $\ker \phi_\beta$, giving rise to a new kernel persistence module, with persistence diagram denoted by $\mathrm{Dgm}(\ker \phi)$. Similarly, we obtain a cokernel persistence module, with diagram $\mathrm{Dgm}(\mathrm{cok}\, \phi)$.

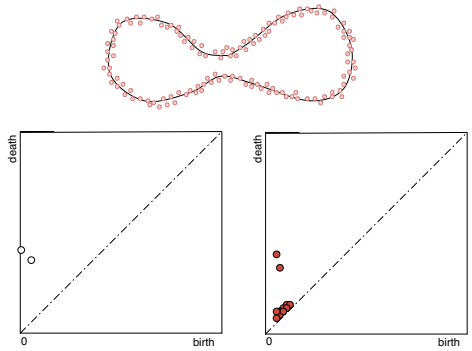

Figure 2: Illustration of a point cloud and its persistence diagram. Top: $\mathbb{X}$ is the curve embedded as shown in the plane and U is the point cloud. Bottom left: the persistence diagram $\mathrm{Dgm}_1(d_\mathbb{X})$; Bottom right: the persistence diagram $\mathrm{Dgm}_1(d_U)$.

Homology

Our main examples of persistence modules all come from homology groups, either absolute or relative, and the various maps between them. Homology persistence modules can arise from families of topological spaces $\{\mathbb{X}_\alpha\}$, along with inclusions $\mathbb{X}_\alpha \hookrightarrow \mathbb{X}_\beta$ for all $\alpha \le \beta$. Whenever we have such a family, the inclusions induce maps $\mathsf{H}_j(\mathbb{X}_\alpha) \to \mathsf{H}_j(\mathbb{X}_\beta)$, for each homological dimension $j \ge 0$, and hence we have persistence modules for each j. Defining $\mathsf{H}(\mathbb{X}_\alpha) = \bigoplus_j \mathsf{H}_j(\mathbb{X}_\alpha)$ and taking direct sums of maps in the obvious way, will also give one large direct-sum persistence module $\{\mathsf{H}(\mathbb{X}_\alpha)\}$.

Distance Functions Here, the families of topological spaces will be produced by the sublevel sets of distance functions. Given a topological space $\mathbb{X}$ embedded in some Euclidean space $\mathbb{R}^N$, we define $d_\mathbb{X}$ as the distance function which maps each point in the ambient space to the distance from its closest point in $\mathbb{X}$. More formally, for each $y \in \mathbb{R}^N$, $d_\mathbb{X}(y) = \inf_{x \in \mathbb{X}} \mathrm{dist}\,(x, y)$. We let $\mathbb{X}_\alpha$ denote the sublevel set $d^{-1}_\mathbb{X}[0, \alpha]$; each sublevel set should be thought of as a thickening of $\mathbb{X}$ within the ambient space. Increasing the thickening parameter produces a growing family of sublevel sets, giving rise to the persistence module $\{\mathsf{H}(\mathbb{X}_\alpha)\}_{\alpha \in \mathbb{R}_{\ge 0}}$; we denote the persistence diagram of this module by $\mathrm{Dgm}(d_\mathbb{X})$ and use $\mathrm{Dgm}_j(d_\mathbb{X})$ for the diagrams of the individual modules for each homological dimension j.

In Figure 2, we see an example of such an $\mathbb{X}$ embedded in the plane, along with the persistence diagram $\mathrm{Dgm}_1(d_\mathbb{X})$. We also have the persistence diagram $\mathrm{Dgm}_1(d_U)$, where U is a dense point sample of $\mathbb{X}$. Note that the two diagrams are quite close in bottleneck distance. Indeed, the difference between the two diagrams will always be upper-bounded by

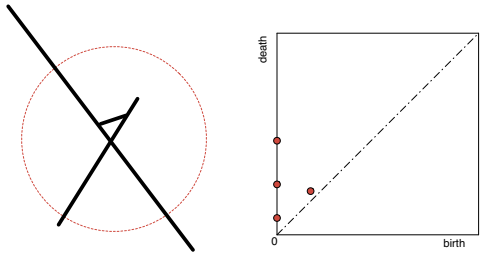

Figure 3: Left: The space $\mathbb{X}$ is in solid line and the closed ball B has dotted boundary. Right: the persistence diagram for the module $\{\mathsf{H}_1(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)\}$.

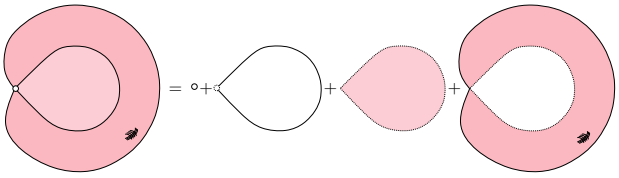

Figure 4: The coarsest stratification of a pinched torus with a spanning disc stretched across the hole.

the Hausdorff distance between the space and its sample.

Persistence modules of relative homology groups also arise from families of pairs of spaces, as the next example shows. Referring to the left part of Figure 3, we let $\mathbb{X}$ be the space drawn in solid lines and B the closed ball whose boundary is drawn as a dotted circle. By restricting $d_{\mathbb{X}}$ to B and also to ∂B, we produce pairs of sublevel sets $(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)$. Using the maps induced by the inclusions of pairs, we obtain the persistence module $\{\mathsf{H}(\mathbb{X}_\alpha \cap B, \mathbb{X}_\alpha \cap \partial B)\}_{\alpha \in \mathbb{R}_{\geq 0}}$ of relative homology groups. The persistence diagram, for homological dimension 1, appears on the right half of Figure 3.

Stratified Spaces

We assume that we have a topological space $\mathbb{X}$ embedded in some Euclidean space $\mathbb{R}^N$. A (purely) *d-dimensional stratification of* $\mathbb{X}$ is a decreasing sequence of closed subspaces

$$\mathbb{X} = \mathbb{X}_d \supseteq \mathbb{X}_{d-1} \supseteq \ldots \mathbb{X}_0 \supseteq \mathbb{X}_{-1} = \emptyset,$$

such that for each i, the i-dimensional *stratum* $\mathbb{S}_i = \mathbb{X}_i - \mathbb{X}_{i-1}$ is a (possibly empty) i-manifold. The connected components of $\mathbb{S}_i$ are called i-dimensional *pieces*. This is illustrated in Figure 4, where the space $\mathbb{X}$ is a pinched torus with a spanning disc stretched across the hole. One usually also imposes a requirement to ensure that the various pieces fit together uniformly. We refer to (Hughes and Weinberger 2000) for precise definitions. Loosely speaking, a stratification is a decomposition of $\mathbb{X}$ into strata such that any two points belonging to the same stratum have similar local structure.

Local Homology and Homology Stratifications Recall ((Munkres 1984)) that the local homology groups of a space $\mathbb{X}$ at a point $x \in \mathbb{X}$ are the groups $\mathsf{H}_i(\mathbb{X}, \mathbb{X} - x)$ in each homological dimension i. If $\mathbb{X}$ happens to be a d-manifold, or if x is simply a point in the top-dimensional stratum of a d-dimensional stratification, then these groups are rank one in dimension d and trivial in all other dimensions. On the other hand, the local homology groups for lower-stratum points can be more interesting; for example if x is the crossing point in Figure 5, then $\mathsf{H}_1(\mathbb{X}, \mathbb{X} - x)$ has rank three.

If x and y are close enough points in a particular piece of the same stratum, then there is a natural isomorphism between their local homology groups $\mathsf{H}(\mathbb{X}, \mathbb{X} - x) \cong \mathsf{H}(\mathbb{X}, \mathbb{X} - y)$, which can be understood in the following manner. Taking a small enough radius r and using excision, we see that the two local homology groups in question are in fact just $\mathsf{H}(\mathbb{X} \cap B_r(x), \mathbb{X} \cap \partial B_r(x))$ and $\mathsf{H}(\mathbb{X} \cap B_r(y), \mathbb{X} \cap \partial B_r(y))$. Both of these groups will then map, via intersection of chains, isomorphically into the group $\mathsf{H}(\mathbb{X} \cap B_r(x) \cap B_r(y), \partial(B_r(x) \cap B_r(y))$, and the isomorphism above is then derived from these two maps. See the points in Figure 5 for an illustration of this idea.

In (Rourke and Sanderson 1999), the authors define the concept of a homology stratification of a space $\mathbb{X}$. Briefly, they require a decomposition of $\mathbb{X}$ into pieces such that the locally homology groups are locally constant across each piece; more precisely, that the maps discussed above be isomorphisms for each pair of close enough points in each piece. This is interesting because in computations we will not be able to distinguish anything finer.

3 Topological Inference Theorem

From the discussion above, it is easy to see that any stratification of a topological space will also be a homology stratification. The converse is unfortunately false. However, we can build a useful analytical tool based on the contrapositive: given two points in a point cloud, we can hope to state, based on their local homology groups and the maps between them, that the two points should not be placed in the same piece of any stratification. To do this, we first adapt the definition of these local homology maps into a more multi-scale and robust framework. More specifically, we introduce a radius parameter r and a notion of local equivalence, $\sim_r$, which allows us to group the points of $\mathbb{X}$, as well as of the ambient space, into strata at this radius scale. We then give the main result of this section: topological conditions under which the point cloud U can be used to infer the strata at different radius scales.

Local Equivalence

We assume that we are given some topological space $\mathbb{X}$ embedded in some Euclidean space in $\mathbb{R}^N$. For each radius $r \geq 0$, and for each pair of points $p, q \in \mathbb{R}^N$, we define the following homology map $\phi^{\mathbb{X}}(p, q, r)$:

$$\begin{aligned} &\mathsf{H}(\mathbb{X} \cap B_r(p), \mathbb{X} \cap \partial B_r(p)) \\ \rightarrow &\mathsf{H}(\mathbb{X} \cap B_r(p) \cap B_r(q), \mathbb{X} \cap \partial(B_r(p) \cap B_r(q))). \end{aligned} \quad (1)$$

Intuitively, this map can be understood as taking a chain, throwing away the parts that lie outside the smaller range, and then modding out the new boundary. Alternatively, one may think of it as being induced by a combination of inclusion and excision.

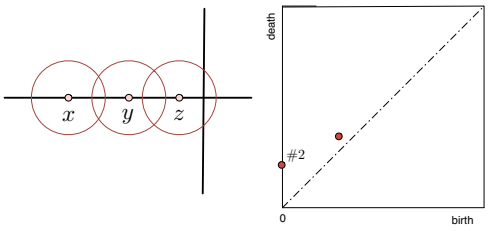

Figure 5: Left: $x \sim_r y$, $y \nsim_r z$. Right: the 1-dim persistence diagram, for the kernel of the map going from the z ball into its intersection with the y ball. A number, i.e., $\#2$, labeling a point in the persistence diagram indicates its multiplicity.

Using these maps, we impose an equivalence relation on $\mathbb{R}^N$.

Definition 3.1 (Local equivalence) *Two points x and y are said to have equivalent local structure at radius r, denoted $x \sim_r y$, iff there exists a chain of points $x = x_0, x_1, \ldots, x_m = y$ from $\mathbb{X}$ such that, for each $1 \leq i \leq m$, the maps $\phi^{\mathbb{X}}(x_{i-1}, x_i, r)$ and $\phi^{\mathbb{X}}(x_i, x_{i-1}, r)$ are both isomorphisms.*

In other words, x and y have the same local structure at this radius iff they can be connected by a chain of points which are pairwise close enough and whose local homology groups at radius r map into each other via intersection. Different choices of r will of course lead to different equivalence classes. For example, consider the space $\mathbb{X}$ drawn in the plane as shown in the left half of Figure 5. At the radius drawn, point z is equivalent to the cross point and is not equivalent to either the point x or y. Note that some points from the ambient space will now be considered equivalent to x and y, and some others will be equivalent to z.

On the other hand, a smaller choice of radius would result in all three of x, y, and z belonging to the same equivalence class.

(Co)Kernel Persistence In order to relate the point cloud U to the equivalence relation $\sim_r$, we must first define a multi-scale version of the maps $\phi^{\mathbb{X}}(p, q, r)$; we do so by gradually thickening the space $\mathbb{X}$. Let $d_{\mathbb{X}} : \mathbb{R}^N \to \mathbb{R}$ denote the function which maps each point in the ambient space to the distance from its closest point on $\mathbb{X}$. For each $\alpha \geq 0$, we define $\mathbb{X}_\alpha = d_{\mathbb{X}}^{-1}[0, \alpha]$. For each p, q, and r, we will consider the intersection map $\phi^{\mathbb{X}}_\alpha(p, q, r)$, which is defined by substituting $\mathbb{X}_\alpha$ for $\mathbb{X}$ in (1). Note of course that $\phi^{\mathbb{X}}(p, q, r) = \phi^{\mathbb{X}}_0(p, q, r)$.

For the moment, we fix a choice of p, q, and r, and we use the following shorthand: $B^{\mathbb{X}}_p(\alpha) = \mathbb{X}_\alpha \cap B_r(p)$, $\partial B^{\mathbb{X}}_p(\alpha) = \mathbb{X}_\alpha \cap \partial B_r(p)$, $B^{\mathbb{X}}_{pq}(\alpha) = \mathbb{X}_\alpha \cap B_r(p) \cap B_r(q)$, $\partial B^{\mathbb{X}}_{pq}(\alpha) = \mathbb{X}_\alpha \cap \partial(B_r(p) \cap B_r(q))$, and we also often write $B^{\mathbb{X}}_p = B^{\mathbb{X}}_p(0)$ and $B^{\mathbb{X}}_{pq} = B^{\mathbb{X}}_{pq}(0)$. By replacing $\mathbb{X}$ with U in this shorthand, we also write $B^U_p(\alpha) = U_\alpha \cap B_r(p)$, and so forth.

For any pair of non-negative real values $\alpha \leq \beta$ the inclusion $\mathbb{X}_\alpha \hookrightarrow \mathbb{X}_\beta$ gives rise to the following commutative diagram:

$$\begin{array}{ccc} \mathsf{H}(B^{\mathbb{X}}_p(\alpha), \partial B^{\mathbb{X}}_p(\alpha)) & \xrightarrow{\phi^{\mathbb{X}}_\alpha} & \mathsf{H}(B^{\mathbb{X}}_{pq}(\alpha), \partial B^{\mathbb{X}}_{pq}(\alpha)) \\ \downarrow & & \downarrow \\ \mathsf{H}(B^{\mathbb{X}}_p(\beta), \partial B^{\mathbb{X}}_p(\beta)) & \xrightarrow{\phi^{\mathbb{X}}_\beta} & \mathsf{H}(B^{\mathbb{X}}_{pq}(\beta), \partial B^{\mathbb{X}}_{pq}(\beta)) \end{array} \quad (2)$$

Hence there are maps $\ker \phi^{\mathbb{X}}_\alpha \to \ker \phi^{\mathbb{X}}_\beta$ and $\operatorname{cok} \phi^{\mathbb{X}}_\alpha \to \operatorname{cok} \phi^{\mathbb{X}}_\beta$. Allowing α to increase from 0 to ∞ gives rise to two persistence modules, $\{\ker \phi^{\mathbb{X}}_\alpha\}$ and $\{\operatorname{cok} \phi^{\mathbb{X}}_\alpha\}$, with diagrams $\mathrm{Dgm}(\ker \phi^{\mathbb{X}})$ and $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})$. Recall that a homomorphism is an isomorphism iff its kernel and cokernel are both zero. In our context then, the map $\phi^{\mathbb{X}}$ is an isomorphism iff neither $\mathrm{Dgm}(\ker \phi^{\mathbb{X}})$ nor $\mathrm{Dgm}(\operatorname{cok} \phi^{\mathbb{X}})$ contain any points on the y-axis above 0.

Examples As shown in the left part of Figure 5, x, y and z are points sampled from a cross embedded in the plane. Taking r as drawn, we note that the right part of the figure displays $\mathrm{Dgm}_1(\ker \phi^{\mathbb{X}})$, where $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(z, y, r)$; we now explain this diagram in some detail. The group $\mathsf{H}_1(B^{\mathbb{X}}_z, \partial B^{\mathbb{X}}_z)$ has rank three; as a possible basis we might take the three classes represented by the horizontal line across the ball, the vertical line across the ball, and the two short segments defining the northeast-facing right angle. Under the intersection map $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}_0$, the first of these classes maps to the generator of $\mathsf{H}_1(B^{\mathbb{X}}_{zy}, \partial B^{\mathbb{X}}_{zy})$, while the other two map to zero. Hence $\ker \phi^{\mathbb{X}}_0$ has rank two. Both classes in this kernel eventually die, one at the α value which fills in the northeast corner of the larger ball, and the other at the α value which fills in the entire right half; these two values are the same here due to symmetry in the picture. At this value, the map $\phi^{\mathbb{X}}_\alpha$ is an isomorphism and it remains so until the intersection of the two balls fills in completely. This gives birth to a new kernel class which subsequently dies when the larger ball finally fills in. The diagram $\mathrm{Dgm}_1(\ker \phi^{\mathbb{X}})$ thus contains three points; the leftmost two show that the map $\phi^{\mathbb{X}}$ is not an isomorphism.

Inference Theorem

Given a point cloud U sampled from $\mathbb{X}$ consider the following question: for a radius r, how can we infer whether or not any given pair of points in U has the same local structure at this radius? In this subsection, we prove a theorem which describes the circumstances under which we can make the above inference. Naturally, any inference will require that we use U to judge whether or not the maps $\phi^{\mathbb{X}}(p, q, r)$ are isomorphisms. The basic idea is that if U is a dense enough sample of $\mathbb{X}$, then the (co)kernel diagrams defined by U will be good enough approximations of the diagrams defined by $\mathbb{X}$.

(Co)Kernel Stability Again we fix p, q, and r, and write $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$. For each $\alpha \geq 0$, we let $U_\alpha = d_U^{-1}[0, \alpha]$. We consider $\phi^U_\alpha = \phi^U_\alpha(p, q, r)$, defined by replacing $\mathbb{X}$ with U_α in (1). Running α from 0 to ∞, we obtain two more persistence modules, $\{\ker \phi^U_\alpha\}$ and $\{\operatorname{cok} \phi^U_\alpha\}$, with diagrams $\mathrm{Dgm}(\ker \phi^U)$ and $\mathrm{Dgm}(\operatorname{cok} \phi^U)$.

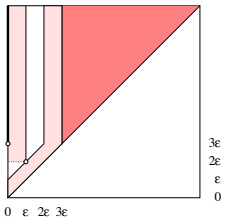

Figure 6: The point in the $\mathbb{X}$-diagrams lie either along the solid black line or in the darkly shaded region. Adding the lightly shaded regions, we get the region of possible points in the U-diagrams.

If U is a dense enough sample of $\mathbb{X}$, then the (co)kernel diagrams defined by U will be good approximations of the diagrams defined by $\mathbb{X}$. More precisely, we have the following consequence of the diagram stability result in (Chazal et al. 2009a):

Theorem 3.1 ((Co)Kernel Diagram Stability) *The bottleneck distances between the (co)kernel diagrams of ϕ^U and $\phi^{\mathbb{X}}$ are upper-bounded by the Hausdorff distance between U and $\mathbb{X}$:*

$$d_B(\mathrm{Dgm}(\ker \phi^U), \mathrm{Dgm}(\ker \phi^{\mathbb{X}})) \leq d_H(U, \mathbb{X}),$$
$$d_B(\mathrm{Dgm}(\mathrm{cok}\, \phi^U), \mathrm{Dgm}(\mathrm{cok}\, \phi^{\mathbb{X}})) \leq d_H(U, \mathbb{X}).$$

Main Inference Result We now suppose that we have a point sample U of a space $\mathbb{X}$, where the Hausdorff distance between the two is no more than some ϵ; in this case, we call U an *ε-approximation* of $\mathbb{X}$. Given two points $p, q \in U$ and a fixed radius r, we set $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$, and we wish to determine whether or not $\phi^{\mathbb{X}}$ is an isomorphism. Since we only have access to the point sample U, we instead compute the diagrams $\mathrm{Dgm}(\ker \phi^U)$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^U)$. The main Theorem of this section, Theorem 3.2, gives conditions under which these diagrams enable us to answer the isomorphism question for $\phi^{\mathbb{X}}$. To state the theorem we first need some more definitions.

Given any persistence diagram $\mathcal{D}$, which we recall is a multi-set of points in the extended plane, and two positive real numbers $a < b$, we let $\mathcal{D}(a, b)$ denote the intersection of $\mathcal{D}$ with the portion of the extended plane which lies above $y = b$ and to the left of $x = a$; note that these points correspond to classes which are born no later than a and die no earlier than b.

For a fixed choice of p, q, r, we consider the following two persistence modules: $\{\mathsf{H}(B_p^{\mathbb{X}}(\alpha), \partial B_p^{\mathbb{X}})\}$ and $\{\mathsf{H}(B_{pq}^{\mathbb{X}}(\alpha), \partial B_{pq}^{\mathbb{X}})\}$. We let $\sigma(p, r)$ and $\sigma(p, q, r)$ denote their respective feature sizes and then set $\rho(p, q, r)$ to their minimum.

We now give the main theorem of this section, which states that we can use U to decide whether or not $\phi^{\mathbb{X}}(p, q, r)$ is an isomorphism as long as $\rho(p, q, r)$ is large enough relative to the sampling density.

Theorem 3.2 (Topological Inference Theorem) *Suppose that we have an ϵ-sample U from $\mathbb{X}$. Then for each pair of points $p, q \in \mathbb{R}^N$ such that $\rho = \rho(p, q, r) \geq 3\epsilon$, the map $\phi^{\mathbb{X}} = \phi^{\mathbb{X}}(p, q, r)$ is an isomorphism iff*

$$\mathrm{Dgm}(\ker \phi^U)(\epsilon, 2\epsilon) \cup \mathrm{Dgm}(\mathrm{cok}\, \phi^U)(\epsilon, 2\epsilon) = \emptyset.$$

This is illustrated in Figure 6. The proof follows, after some work, from the persistence diagram stability results in (Chazal et al. 2009a).

Examples Here we give two examples illustrating the topological inference theorem. For the first example, as shown in Figure 7, suppose we have $\mathbb{X}$ in the top left and we take the points p and q and r as drawn; in this case, one can show that $\rho(p, q, r) = 8.5$, which here is the distance between the line segment and the boundary of the intersection of the two r-balls. First we compute the (co)kernel persistence diagrams for $\phi^{\mathbb{X}}$, showing only the kernel diagram in the top right. Since the y-axis of this diagram is free of any points (and the same holds for the un-drawn cokernel diagram), p and q have the same local structure at this radius level. On the other hand, suppose that we have an ϵ-sample U of $\mathbb{X}$, with $\epsilon = 2.8 < \rho/3$, as drawn in the bottom left. We can compute the analogous U-diagrams, with the kernel diagram drawn in the bottom right. Noting that the rectangle defined by $(\epsilon, 2\epsilon)$ in the diagram is indeed empty, and that the same holds for the cokernel diagrams, we can apply Theorem 3.2 to infer that the points have the same local structure at radius level r.

For the second example, as shown in Figure 8, suppose $\mathbb{X}$ is the cross on the top left with p, q, r as drawn. Then p and q are locally different at this radius level, as shown by the presence of two points on the y-axis of the kernel persistence diagram. In the bottom left, we show an ϵ-sample U of $\mathbb{X}$, with $3\epsilon < \rho(p, q, r)$. Note that the kernel diagram for ϕ^U does indeed have two points in the relevant rectangle, therefore indicating different local structure.

4 Probabilistic Inference Theorem

The topological inference of Section 3 states conditions under which the point sample U can be used to infer stratification properties of the space $\mathbb{X}$. The basic condition is that the Hausdorff distance between the two must be small. In this section we describe two probabilistic models for generating the point sample U, and we provide an estimate of how large this point sample should be to infer stratification properties of the space $\mathbb{X}$ with a quantified measure of confidence. More specifically, we provide a local estimate, based on $\rho(p, q, r)$ and $\rho(q, p, r)$, of how many sample points are needed to infer the local relationship at radius level r between two fixed points p and q; this same theorem can be used to give a global estimate of the number of points needed for inference between any pair of points whose ρ-values are above some fixed low threshold.

Sampling Strategies

We assume $\mathbb{X}$ to be compact. Since the stratified space $\mathbb{X}$ can contain singularities and maximal strata of varying dimensions, some care is required in the sampling design. Consider for example a sheet of area one, punctured by a line of

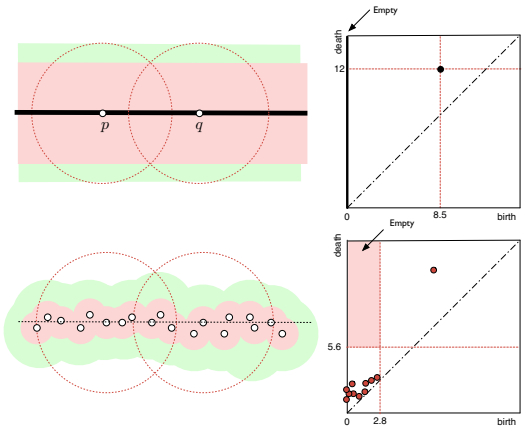

Figure 7: Kernel persistence diagram of two local equivalent points, given $\mathbb{X}$ (top) and given U (bottom).

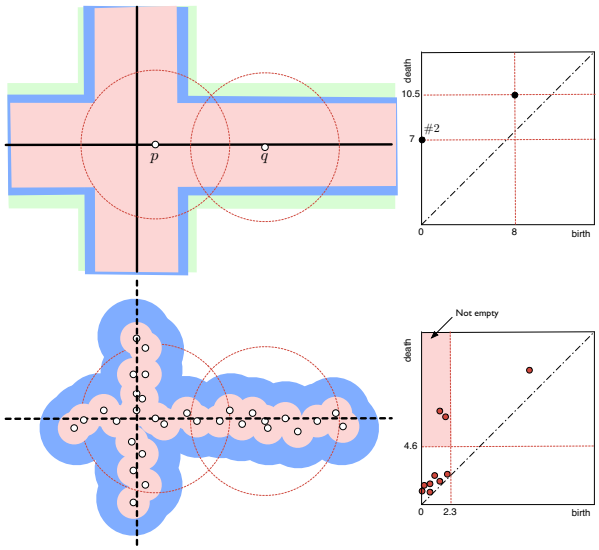

Figure 8: Kernel persistence diagram of two points that are not locally equivalent, given $\mathbb{X}$ (top) and given U (bottom). A number, i.e., $\#2$, labeling a point in the persistence diagram indicates its multiplicity.

length one. In this case, sampling from a naively constructed uniform measure on this space would result in no points being drawn from the line. This same issue arose and was dealt with in (Niyogi, Smale, and Weinberger 2008b), although in a slightly different approach than we will develop.

The first sampling strategy is to remove the problems of singularities and varying dimension by replacing $\mathbb{X}$ by a slightly thickened version $\mathbb{X} \equiv \mathbb{X}_\delta$. We assume that $\mathbb{X}$ is embedded in $\mathbb{R}^N$ for some N. This new space is a smooth manifold with boundary and our point sample is a set of n points drawn identically and independently from the uniform measure $\mu(\mathbb{X})$ on $\mathbb{X}$, $U = \{x_1, ..., .x_n\} \overset{iid}{\sim} \mu(\mathbb{X})$. This model can be thought of as placing an appropriate measure on the highest dimensional strata to ensure that lower dimensional strata will be sampled from. We call this model M_1.

The second sampling strategy is to deal with the problem of varying dimensions using a mixture model. In the example of the sheet and line, a uniform measure would be placed on the sheet, while another uniform measure would be placed on the line, and a mixture probability is placed on the two measures; for example, each measure could be drawn with probability $1/2$. We now formalize this approach. Consider each (non-empty) i-dimensional stratum $\mathbb{S}_i = \mathbb{X}_i - \mathbb{X}_{i-1}$ of $\mathbb{X}$. All strata that are included in the closure of some higher-dimensional strata, in other words all non-maximal strata, are not considered in the model. A uniform measure is assigned to the closure of each maximal stratum, $\mu_i(\mathbb{S}_i)$, this is possible since each such closure is compact. We assume a finite number of maximal strata K and assign to the closure of each such stratum a probability $p_i = 1/K$. This implies the following density, $f(x) = \frac{1}{K}\sum_{j=1}^{K} \nu_i(X = x)$, where ν_i is the density corresponding to measure μ_i. The point sample is generated from the following model: $U = \{x_1, ..., .x_n\} \overset{iid}{\sim} f(x)$. We call this model M_2.

The first model replaces a stratified space with its thickened version, which enables us to place a uniform measure on the thickened space. Although this replacement makes it convenient for sampling, it does not sample directly from the actual space. The second model samples from the actual space, however the sample is not uniform on $\mathbb{X}$ with respect to Lebesgue measure.

Lower bounds on the sample size of the point cloud

Our main theorem is the probabilistic analogue of Theorem 3.2. An immediate consequence of this theorem is that, for two points $p, q \in U$, we can infer with probability at least $1 - \xi$ whether p and q are locally equivalent, $p \sim_r q$. The confidence level $1 - \xi$ will be a monotonic function of the size of the point sample.

The theorem involves a parameter $v(\rho)$, for each positive ρ, which is based on the volume of the intersection of ρ-balls with $\mathbb{X}$. First we note that each maximal stratum of $\mathbb{X}$ comes with its own notion of volume: in the plane punctured by a line example, we measure volume in the plane and in the line as area and length, respectively. The volume $\mathrm{vol}\,(\mathbb{Y})$ of any subspace $\mathbb{Y}$ of $\mathbb{X}$ is the sum of the volumes of the intersections of $\mathbb{Y}$ with each maximal stratum. For $\rho > 0$, we define

$$v(\rho) = \inf_{x \in \mathbb{X}} \frac{\mathrm{vol}\,(B_{\rho/24}(x) \cap \mathbb{X})}{\mathrm{vol}\,(\mathbb{X})} \tag{3}$$

We can then state:

Theorem 4.1 (Local Probabilistic Sampling Theorem) *Let $U = \{x_1, x_2, ..., x_n\}$ be drawn from either model M_1 or M_2. Fix a pair of points $p, q \in \mathbb{R}^N$ and a positive radius r, and put $\rho = \min\{\rho(p, q, r), \rho(q, p, r)\}$. If*

$$n \geq \frac{1}{v(\rho)}\left(\log\frac{1}{v(\rho)} + \log\frac{1}{\xi}\right),$$

then, with probability greater than $1-\xi$ we can correctly infer whether or not $\phi^{\mathbb{X}}(p,q,r)$ and $\phi^{\mathbb{X}}(q,p,r)$ are both isomorphisms.

To extend the above theorem to a more global result, one can pick a positive ρ and radius r, and consider the set of all pairs of points (p,q) such that $\rho \leq \min\{\rho(p,q,r), \rho(q,p,r\}$. Applying Theorem 4.1 uniformly to all pairs of points will give the minimum number of sample points needed to settle the isomorphism question for all of the intersection maps between all pairs.

5 Algorithm

The theorems in the last sections give conditions under which a point cloud U, sampled from a stratified space $\mathbb{X}$, can be used to infer the local equivalences between points on $\mathbb{X}$ and its surrounding ambient space. We now describe clustering U-points themselves into strata. We imagine that we are given the following input: a point cloud U sampled from some stratified space $\mathbb{X}$, and a fixed radius r. We make the assumption that $d_H(U,\mathbb{X}) \leq \epsilon \leq \frac{\rho_{\min}}{3}$, where $\rho_{\min}$ is the minimum of $\rho(p,q,r)$ for all pairs $(p,q) \in U \times U$. Later we discuss the consequences when this assumption does not hold and a possible solution.

We build a graph where each node in the graph corresponds uniquely to a point from U. Two points $p,q \in U$ (where $||p-q|| \leq 2r$) are connected by an edge iff both $\phi^{\mathbb{X}}(p,q,r)$ and $\phi^{\mathbb{X}}(q,p,r)$ are isomorphisms, equivalently iff $\mathrm{Dgm}(\ker \phi^U)(\epsilon, 2\epsilon)$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^U)(\epsilon, 2\epsilon)$ are empty. The connected components of the resulting graph are our clusters. Note that the connectivity of the graph is encoded by a weight matrix, and our clustering strategy is based on a $0/1$-weight assignment.

A crucial subroutine in the clustering algorithm is the computation of the diagrams $\mathrm{Dgm}(\ker \phi^U)$ and $\mathrm{Dgm}(\mathrm{cok}\, \phi^U)$. Basically we use simplicial complexes constructed from Delaunay triangulation and the (co)kernel persistence algorithm described in (Cohen-Steiner et al. 2009). It is quite complicated, we defer all details to our technical report (Bendich, Mukherjee, and Wang 2010).

Robustness of clustering Two types of errors in the clustering can occur: false positives where the algorithm connects points that should not be connected and false negatives where points that should be connected are not. The current algorithm we state is somewhat brittle with respect to both false positives as well as false negatives. The false positives are driven by the condition in Theorem 3.2 that $\rho_{\min} < 3\epsilon$, so if the point cloud is not sampled fine enough we can get incorrect positive isomorphisms and therefore incorrect edges in the graph. If we use transitive closure to define the connected components this can be very damaging in practice since a false edge can collapse disjoint components into one large cluster. If we replace transitive closure with a spectral clustering approach we will have a more robust clustering or assignments. It is natural to think of the $0/1$-weight assignment on pairs of points $p,q \in U$ as an association matrix $\mathbf{W}$. Given this association matrix we can use spectral clustering to obtain a robust partition of the points (Meilă and Shi 2001; Kannan, Vempala, and Veta 2000).

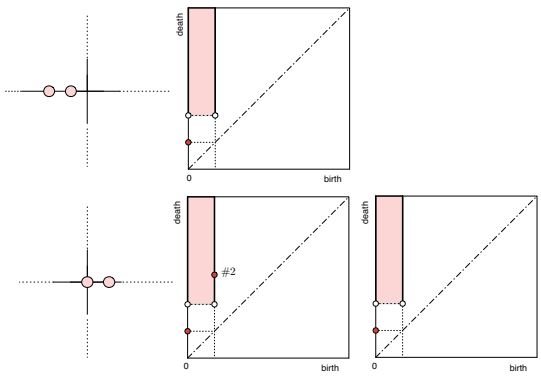

Figure 9: Top: both points are from 1-strata. Bottom: one point from 0-strata, one point from 1-strata. Left part shows the locations of the points. Right part shows the ker/cok persistence diagram of two points respectively, if the diagrams are the same, only one is shown.

6 Simulations

We use a simulation on simple synthetic data with points sampled from grids to illustrate how the algorithm performs. In these simulations we assume we know ε, and we run our algorithm for $0 \leq \alpha \leq 2\varepsilon$. The three data sets tested are: points sampled from a cross; points sampled from a plane intersecting a line; points sampled from two intersecting planes. We use the following result to demonstrate that the inference on local structure, at least for these very simple examples, is correct. As shown in Figure 9 top, if two points are locally equivalent, their corresponding ker/cok persistence diagrams contain the empty quadrant prescribed by our theorems, while in Figure 9 bottom, the diagrams associated to two non-equivalent points do not contain such empty quadrants. Similar results are shown for the other data sets in Figure 10 and 11, where a number labeling a point in the persistence diagram indicates its multiplicity.

7 Discussion

We would like to make clear that we consider the algorithm in this paper a first step and several issues both statistical and computational can be improved upon. We state a few extensions of interest here. The algorithm to compute the (co)kernel diagrams from the thickened point cloud should be quite slow when the dimensionality of the ambient space is high due to the runtime complexity of Delaunay triangulation. Is there a faster way, for example, using Rips or Witness complexes (de Silva and Carlsson 2004)? Currently we use a graph with $0/1$ weights based on the local equivalence between two points. Extending this idea to assign fractional weights between points is appealing as it suggests a more continuous metric for local equivalence. This may also allow for greater robustness when using spectral methods to assign points to strata.

Acknowledgments

All the authors would like to thank Herbert Edelsbrunner and John Harer for suggestions. PB would like to thank

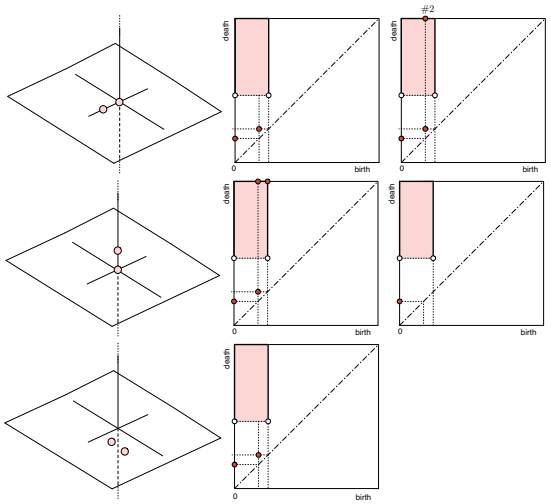

Figure 10: Top: one point from 0-strata, one point from 2-strata. Middle: one point from 0-strata, one from 1-strata. Bottom: both points are from 2-strata.

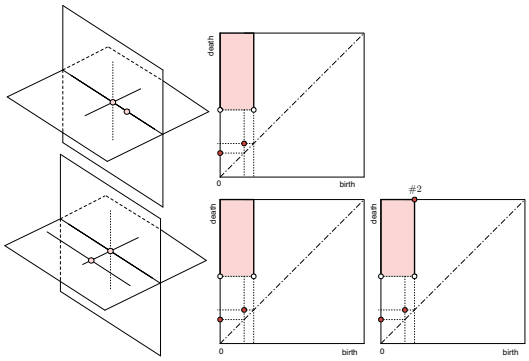

Figure 11: Top: both points from 1-strata. Middle: one point from 1-strata, one from 2-strata.

David Cohen-Steiner and Dmitrity Morozov for helpful discussion, BW would like to thank Yuriy Mileyko for insightful discussion and Mikael Vejdemo-Johansson for providing the Plex package, and SM would like to thank Shmuel Weinberger for useful comments. SM and BW would like to acknowledge the support of NIH Grants R01 CA123175-01A1 and P50 GM 081883, and SM would like to acknowledge the support of NSF Grant DMS-07-32260. PB thanks the Computer Science Department at Duke University for hosting him during the Spring semester of 2010.

References

Bendich, P.; Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J.; and Morozov, D. 2007. Inferring local homology from sampled stratified spaces. In *Proceedings 48th Annual IEEE Symposium on Foundations of Computer Science*, 536–546.

Bendich, P.; Mukherjee, S.; and Wang, B. 2010. Towards stratification learning through homology inference. Manuscript, http://arxiv.org/abs/1008.3572.

Chazal, F.; Cohen-Steiner, D.; Glisse, M.; Guibas, L. J.; and Oudot, S. Y. 2009a. Proximity of persistence modules and their diagrams. In *Proceedings 25th Annual Symposium on Computational Geometry*, 237–246.

Chazal, F.; Guibas, L. J.; Oudot, S. Y.; and Skraba, P. 2009b. Analysis of scalar fields over point cloud data. In *Proceedings 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1021–1030.

Chazal, F.; Cohen-Steiner, D.; and Lieutier, A. 2009. A sampling theory for compact sets in euclidean space. *Discrete and Computational Geometry* 41:461–479.

Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J.; and Morozov, D. 2009. Persistence homology for kernels, images and cokernels. In *Proceedings 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1011–1020.

Cohen-Steiner, D.; Edelsbrunner, H.; and Harer, J. 2007. Stability of persistence diagrams. *Discrete and Computational Geometry* **37**:103–120.

de Silva, V., and Carlsson, G. 2004. Topological estimation using witness complexes. *Symposium on Point-Based Graphics* 157–166.

Edelsbrunner, H., and Harer, J. 2010. *Computational Topology: An Introduction*. Providence, RI, USA: American Mathematical Society.

Goresky, M., and MacPherson, R. 1988. *Stratified Morse Theory*. New York, NY, USA: Springer-Verlag.

Haro, G.; Randall, G.; and Sapiro, G. 2007. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *Advances in Neural Information Processing Systems* 19:553–560.

Hughes, B., and Weinberger, S. 2000. Surgery and stratified spaces. *Surveys on Surgery Theory* 311–342.

Kannan, R.; Vempala, S.; and Veta, A. 2000. On clusteringsgood, bad and spectral. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 367.

Lerman, G., and Zhang, T. 2010. Probabilistic recovery of multiple subspaces in point clouds by geometric lp minimization. Manuscript.

Meilă, M., and Shi, J. 2001. Learning segmentation by random walks. *Advances in Neural Information Processing Systems* 13:873—879.

Munkres, J. R. 1984. *Elements of algebraic topology*. Redwood City, CA, USA: Addison-Wesley.

Niyogi, P.; Smale, S.; and Weinberger, S. 2008a. Finding the homology of submanifolds with high confidence from random samples. *Discrete Computational Geometry* 39:419–441.

Niyogi, P.; Smale, S.; and Weinberger, S. 2008b. A topological view of unsupervised a topological view of unsupervised learning from noisy data. Manuscript.

Rourke, C., and Sanderson, B. 1999. Homology stratifications and intersection homology. *Geometry and Topology Monographs* 2:455–472.

Singh, G.; Mémoli, F.; and Carlsson, G. 2007. Topological methods for the analysis of high dimensional data sets and 3D object recognition. *Eurographics Symposium on Point-Based Graphics* 91–100.

Vidal, R.; Ma, Y.; and Sastry, S. 2005. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27:1945 – 1959.

Weinberger, S. 1994. *The topological classification of stratified spaces*. Chicago, IL, USA: University of Chicago Press.