Are the data hollow?
Leonid Pekelis, Susan Holmes
Statistics Department, Sequoia Hall, Stanford University, CA 94305
December 14, 2009

Abstract
This paper presents statistical approaches to testing whether data are spherical in the topological sense, i.e., they lie on a closed manifold with a hollow interior. We characterize the sphere by its Betti numbers and use the computational topology program PLEX [12] based on simplicial complexes to calculate the data’s Betti numbers.
We use the parametric bootstrap approach to test the power of the test for several alternatives including the full sphere or ellipsoid filling points in three dimensions.

1 Introduction
High dimensional data sometimes lie on a lower dimensional manifold, this has been extensively explored by the manifold learning community for examples see [17], [1] or [13].
However in some applications it is important to know whether the manifold is closed as in the simple case of cell cycle data where gene expressions in very high dimensions can be projected onto a closed curve, see [7] for the details of the cell cycle application.
It can be important to see if data has a hollow structure, even in data that are originally known to come from a sphere, such as geological measurements. However the nature of the data themselves often obscure the fact that the data are hollow, since spherical data will project automatically on the circle in two dimensions. Dynamic graphics such as those proposed by ggobi[15] can provide a way of ‘seeing’ that the data sit on a sphere, but we are interested in a statistical test for such situations.
In other types of data, such as astronomical data some researchers [19] have used statistical methods to detect topological structure, we propose an approach also based on topology aimed at detecting structure with dense surfaces surrounding hollow volumes that can be characterized by a specific pattern of the topological characteristics called Betti numbers[10].

2 Methodology
For the remainder of this article, let \( X \) be a random variable in \( \mathbb{R}^d \) with density function \( f : \mathcal{M} \to [0,1] \), and \( \mathcal{M} \) a manifold, possibly of dimension less than \( d \). The general problem in topological statistics is identifying a good approximation to the space \( \mathcal{M} \) from \( X \), a \( n \times d \) matrix encoding \( n \) possibly noisy draws from \( X \). In this paper, we focus the hypothesis test
\[
H_0 : \mathcal{M} \text{ does not have the topology of the sphere } S^d, d \geq 1
\] (1)
In order to reformulate this null hypothesis into a well defined problem, we use the methods of persistent homology from the field of computational topology. The next section summarizes the theory behind persistent homology and the process of approximating topological properties of \( \mathcal{M} \) from \( X \). This leads to a restatement of (1) in topological terms.
For a more thorough introduction to persistent homology, [2] is an excellent place to start.

2.1 Calculating the Homology of Data
The homology of a space \( \mathcal{M} \) is a set of specific topological characteristics which describe \( \mathcal{M} \). Formally, it is a series of functors that takes \( \mathcal{M} \) to algebraic images. Two homeomorphic spaces necessarily have the same homology, but the converse is not always true. In principle, other functors exist, such as the fundamental group, but for data analysis homology is exclusively examined due to computational tractability. A more general approach falls under algebraic topology. For a thorough treatment see [10].
Calculating homology from \( X \) is accomplished in two principle steps. First, build a simplicial complex, denoted by \( S(X, r) \), where \( r \) is a 1-dimensional parameter often relating to the level sets of a distance function on \( X \). And second, calculate the homology groups, \( H_p(S(X, r)) \), where \( p \) runs over \( \mathbb{Z}_+ \) and denotes dimension. Intuitively, for a well chosen value of \( r \), \( S(X, r) \) is a polyhedral approximation to \( \mathcal{M} \). Thus, we can hope for isomorphisms between their homology groups, \( H_p(\mathcal{M}) \approx H_p(S(X, r)) \).
The structure of \( S(X, r) \) is a collection of vertices, edges, triangles, and their higher dimension equivalents, collectively named simplices, together with information on how various simplices fit together. Imagine a multidimensional
jig saw puzzle. There are many different approaches to constructing $S(\mathcal{X}, r)$ in the literature, including Čech, Viteri-Rips, α-shape, and witness. In this paper we focus on witness complexes, and define them as follows. Discussion of the others may be found in [3].

- Let $L \subset \mathcal{X}$ be a set of vertices chosen at random.
- For each $x \in \mathcal{X}$, define $m(x)$ to be the distance from $x$ to the closest landmark, $l \in L$.
- The edge, $[ab]$, is then in $S(\mathcal{X}, t)$ if there exists a “witness” $x \in \mathcal{X}$ such that $\max\{|a - x|, |b - x|\} \leq t + m(x)$.
- All higher dimensional simplices are in $S(\mathcal{X}, t)$ if and only if all of their edges are.

The biggest advantage of using the witness approach is that the number of simplices in $S(\mathcal{X}, t)$ is bounded by a function of $|L|$, the number of landmarks. Witness complexes are also computationally simple. These advantages quickly become necessary in order to finish computation in a reasonable time frame. The trade off is witness complexes currently lack theoretical tractability, so most results regarding them must be given heuristically. For the sake of notation, we now drop the parameters $\mathcal{X}$ and $r$, letting $S = S(\mathcal{X}, r)$ when there is no ambiguity.

Define the boundary map, $\delta_k : C^k \rightarrow C^{k-1}$, which takes $k$-dimensional simplices to their $(k-1)$-dimensional faces. For example, the boundary of a triangle is its 3 edges. Next, define a $k$-cycle as a group of $k$-dimensional simplices with 0 boundary, mod2. The 3 edges of the triangle collectively have 0 boundary since each of the vertices are counted twice, and is therefore a cycle. Then $H_k(S) \cong \mathbb{Z}_k/B_{k+1}$, the $k$-dimensional cycle cosets created by dividing out those cycles which are boundaries of $(k+1)$-dimensional groups of simplices. For instance, the three edges of a triangle would be collapsed in $H_1$. Thus, $H_k$ represents each “true” $k$-dimensional cycle, or hole, as an abelian group, and its rank, $\beta_k = rk(H_k)$, counts the number of $k$-dimensional holes. These ranks are the Betti-$k$ numbers, with $\beta_0$ counting the number of connected components. The Betti-$k$ numbers of $\mathcal{M}$ are its homology signature and are a way to classify topological spaces. A single point and a solid sphere both have the same homology signature, $(1, 0, \ldots, 1)$, and are homologous; $\mathbb{S}^d$ has homology signature $(1, 0, \ldots, 1, 0, \ldots)$, where the second 1 is in the $d$-th position, and are not homologous for distinct $d$.

At this point, it may be tempting to reformulate our null hypothesis as $H_0(\mathcal{S}) \cong \mathbb{Z}$, but this is not in exact correspondence with (1) as in principle $\mathcal{M}$ may have a non-trivial homology not isomorphic to $\mathbb{Z}^d$, with the distribution of $|T(\mathcal{S})|$ being far from that of $|T_{G_0, \mathcal{S}^d}|$. This event can be interpreted by inclusion in the type 1 error of the original test. For results suggesting (2) is a conservative approximation to (1) see section 3.

The goal of the rest of this article will be to examine the power of the test based on the new definition (2) through simulation. We consider two categories of alternative hypotheses.

In the first category, we assume $X$ to have uniform distribution on $\mathbb{S}^d$ convoluted with varying amounts of perpendicular Gaussian noise.

$$H_0^1 : X \sim U(\mathbb{S}^d) \ast N(0, \sigma^2), \sigma \geq 0.$$  (3)
In the second category of alternatives, $H^2_n$, $X$ is considered to be sampled from the surface of a manifold homeomorphic to $S^d$, but not congruent. In particular, we consider the shapes resulting when two ends of a circle or sphere are pushed towards each other. For an example, see figure 1(c).

The next section describes procedures to generate data according to the alternatives.

2.2 Generating Data from a Manifold

Sampling from a general manifold can be done quite simply especially in the case of surfaces of revolution, see [8] provides a review of the subject with statistical applications. Under $H^3_1$, we are interested in testing the noise tolerance for correctly calculating the homology of a d-sphere. We only demonstrate the procedure for $d = 2$, noting that the cases of $d = 1, 3$ are exactly similar, only less informative. Points are generated on the 2-sphere according to the standard parametrization

$$x = \sin(\theta) \cos(\phi) ; \ y = \sin(\theta) \sin(\phi) ; \ z = \cos(\theta) ; \ \theta \in [0, \pi], \ \phi \in [0, 2\pi].$$

(4)

The parametrization imposes a particular density on $\theta$ and $\phi$ given by its Hausdorff area measure. Section 3.2.5 of Federer (1996) gives a formula for explicitly calculating this density for any manifold $M$, which we reproduce below for the case of $M = S^2$.

$$\int f(g(y))J_2g(y)\lambda^2(dy) = \int f(x)N(g|A, x)\mathcal{H}^1(dx)$$

(5)

where $y = (\theta, \phi), x = (x, y, z), g : [0, \pi] \times [0, 2\pi] \rightarrow S^2$ is the parametrization (4), $f : S^2 \rightarrow [0, 1]$ any specified density on $S^2$, $J_2$ the 2-dimensional Jacobian, $N(f|A, y) = \# \{ x \in A : f(x) = y \}$ is the multiplicity function, and $\mathcal{H}^1(dx)$ the 3-dimensional Hausdorff measure. Intuitively, (5) accounts for stretching in the parametrized space due to the curvature of $S^2$. Drawing from $S^2$ with density $f$ now amounts to sampling from the density $g \circ J_m f(y)$, after normalization, in the parameter space. Letting $f$ be the uniform density, it is easy to show that $f_{S^2}(y) \doteq g \circ J_m f(y) = \dfrac{1}{\pi} \sin(\theta)$. Figure 1 (a) displays a heat map of $f_{S^2}(y)$ on $S^2$.

Finally, normal noise is added to each point by specifying a distribution on the radius, $r \sim N(1, \sigma^2)$, and applying the transformation $h(y) = r * g(y)$. The joint distribution of $r, \theta$ and $\phi$ is given by

$$f_X(r, \theta, \phi) = f_{S^2}(\theta, \phi) f(r) = \dfrac{1}{\sqrt{32\pi^3}} e^{-\dfrac{(r-1)^2}{2\sigma^2}}.$$  

(6)

It is easy to show that the expected squared distance of $X$ from $S^2$ is completely determined by the noise variance, $\sigma$. Also, the Gaussian noise is independent from the angular distribution of $X$. Figure 1 (b) shows a sample $X$ from $X$ superimposed on a mesh of $S^2$ with the distance $\sigma^2$ marked.

![Figure 1](image-url)

Figure 1: Figure (a) is a heat map of uniform area measure on $S^2$. Figure (b) shows a sample from $X (\sigma = 0.1)$ superimposed. Figure (c) shows manifolds parametrized by $f_p$ for $\alpha$ in $\{0.25, 0.5, 1, 3, 5\}$ in increasing order of indentation. Figure (d) gives a sample from (7), with $\alpha = 5$.

The second set of alternatives, $H^2_n$, continuously transform $S^1$ to resemble a 1-sphere less and less, while keeping homology intact. Let

$$f^+_p(\theta; r, \alpha) \doteq \left\{ r \cos(\theta), r \sin(\theta) - \left( e^{-\alpha \cos(\theta)} - e^{-\alpha} \right) \right\} ; \ \theta \in [0, \pi] ; \ d, r \in \mathbb{R}_+,$$

$$f^-_p = f^+_p(\theta - \pi; r, \alpha), \text{ and } f_p(\theta; r, \alpha) = I_{\{ \theta \in [0, \pi] \}} f^+_p(\theta; r, \alpha) - I_{\{ \theta \in [\pi, 2\pi] \}} f^-_p(\theta; r, \alpha).$$

Increasing the parameter $\alpha$ mimics the action of placing 2 fingers on opposite sides of a balloon and attempting to touch them together. Manifolds for representatives $\alpha$’s are pictured in figure 1(c). The corresponding (non-normalized) area measure for $f^+_p$ is

$$f_A(\theta; r, \alpha) = r \left[ \sqrt{1 - 2 \cos(\theta)} A(\theta) + A(\theta)^2 \right], \ A(\theta) = \alpha \sin(2\theta) e^{-\alpha \cos^2(\theta)},$$

(7)
and symmetric for $f_p$. Although we only consider the two dimensional case, higher dimension versions of $f_p$ can be calculated through rotations about a bisecting axis.

Although the area measure (7) is clearly integrable, no simple analytical solution exists and we resort to Markov Chain Monte Carlo (MCMC) methods to sample data. Figure 1(d) shows a sample from distribution function $x$, parametrized by $f_p$, and $\alpha = 5$. The sample appears uniformly distributed despite heavy curvature in the manifold.

For numerical analysis, we consider the cases of $d \in \{1, 2, 3\}$ and $\sigma \in [0, 1]$ for $H^1_\sigma$, corresponding to a noise standard deviation up to the radius of $S^d$, and $d = 2, \alpha \in \{1, 2, \ldots, 5\}$ for $H^2_\alpha$.

3 Results
Our first result suggests that using the distribution of $T$ under the null hypothesis of (2) can be regarded as a conservative estimate of the more general test (1). We generate comparison data from bi-variate Gaussian mixtures, centered at the origin. Each mixture consists of two Gaussians with equal absolute variance, rotated either 45 degrees clockwise or counterclockwise. This results in an “X” shaped distribution, examples of which are given in figure 2(a)-(c), with 95% equi-confidence ellipses drawn for each Gaussian. A 2-sample Anderson-Darling test was then performed to compare the resulting empirical distributions of $T$ when 200 samples of 1000 points each are drawn from the mixed distributions above, and the null $G_0$. The resultant p-values are 0.279, 0.016 and 0 suggesting that while there is not enough evidence to prove $T_{G_0}$ and $T$ under distribution (a) have dissimilar distributions, the distribution of $T$ under either case (b) or (c) is significantly different from that of $T_{G_0}$.

From here, we calculate confidence intervals for the “effective $\alpha$” values of the test (2) in the cases (a), (b), and (c). We define the effective $\alpha$ value, $\tilde{\alpha}$, by the type I error rate under an alternate null hypothesis, using the original critical value. The value $\tilde{\alpha}$ answers the following question: assuming the correct null distribution to use was actually (a), (b), or (c), what significance level are we testing by using a critical value from $G_0$? A confidence interval for $\tilde{\alpha}$ can be readily attained by estimating the $G_0$ critical value from its empirical distribution, $T^\alpha_{G_0}$, and then performing 2000 bootstrap resamples of the statistic $p^B(\alpha) = \hat{P}(T \geq T_{G_0})$. Since it is reasonable to assume $p^B(\alpha)$ may not be variance stabilized, we use a bias-corrected and accelerated (BCa) method to calculate bootstrap confidence intervals. Resulting 90% intervals for $p^B(0.05)$ are [0.025, 0.075], [0.015, 0.055], and [0.000, 0.020]. Hence, among these examples, greater dissimilarity of the test statistic $T$ from $T_{G_0}$, in distribution, corresponds to a lower effective alpha.

![Sample taken from Gaussian mixture distribution: $Z = X_1 + X_2$, where $X_1$ and $X_2$ are bivariate Gaussians of equal absolute variance, rotated 45 degrees. A 95% equi-confidence ellipse is drawn for $X_1$ and $X_2$. Figure (d) shows a sample from $G_0$ for comparison.](image)

The next set of tests examine the noise tolerance of $T_X$ with $X$ distributed according to (6), by sampling from $H^1_\sigma$. The dimensions of $S^d$ considered are $d$ in $\{1, 2, 3\}$ with respective sets of noise parameters $\sigma_1 = \{0.05, 0.10, \ldots, 0.95\}$, $\sigma_2 = \{0.05, 0.10, \ldots, 0.70\}$, and $\sigma_3 = \{0.05, 0.10, \ldots, 0.25\}$. Decreasing set sizes for $\sigma_k$ were taken due to computational considerations. As before we take 200 samples of 1000 points for each dimension and noise level. To estimate changes in power directly, we calculate bootstrap confidence intervals for the power of the empirical critical value, $\hat{P}(T_X \geq T^\alpha_{G_0} | X \sim U(S^d) \ast N_1(0, \sigma^2))$, similar to the procedure above. Descriptive results are given in figure 3(a)-(c). The overall message seems to be that recovering a sphere topology is highly unlikely when sampling noise exceeds 30% of the radius of the manifold. We now turn to examine further why this break down occurs.

One test in this direction is to examine which levels of $\sigma$ are critical in determining the distribution of $T_{X(\sigma)}$. That is, what is the set of $\sigma_k$ such that $T_{\sigma_k-1}$ has a different distribution from $T_{\sigma_k}$? Note we have simplified notation by defining $T_{\sigma} = T_{X(\sigma)}$. Figures 3(d)-(f) show quantile-quantile plots comparing the distributions of $T_{\sigma_k}$ to a standard normal. The results indicate a partition into an active period when noise is below 30 – 35% of the radius of $S^d$, and a tame period for greater noise. The active period is characterized by a monotone shift from negative skew centered at a proportion close to 1, to a symmetric distribution, to a positive skew centered about 0. The tame period is characterized by distributions all similar to that of $T_{G_0}$, represented by the bold, red line in figures 3(d)-(f). There is also some evidence of a dimensionality effect. In the case of $S^3$, embedded in 4 dimensions, the noise breakdown
and partitioning $\sigma$ seem to be closer to .25, although we do not have data on higher noise levels to verify this claim. Two sample Anderson-Darling tests comparing distributions for consecutive $\sigma$’s support these results.

A final series of tests inspects the power of the test (2) against a particular class of diffeomorphisms of $\mathbb{S}^d$, namely those in $H^2_a$, and examines the consistency of the estimator $T_X$ in this class.

Confidence intervals for empirical power are computed through similar bootstrap $BC_0$ intervals. As seen in figure 3(g), power is lost as the coefficient $\alpha$ increases, corresponding to indentation of the circle at both sides. This is expected as indentation would cause bisecting edges to form at smaller distances in the simplicial complex. In other words, complexes not homeomorphic to $\mathbb{S}^1$ are present for a larger fraction of distance parameters observed.

A second conclusion is that $T_X$ is a consistent estimator under $H^2_a$. For figure 3(h), we performed the same power calculations as above, fixing $\beta = .30$, and letting the sample size $n$ vary. As $n$ increases, empirical power also
increases, practically to the same level as no indentation.

Although no theoretical results into the consistency of $T_x$ under $H_2$ are given, an intuitive argument is as follows. Since the manifold in question has finite curvature everywhere, there must be some threshold distance $\epsilon$ such that any edge connecting two points with length less than $\epsilon$ does not bisect the manifold in a way that alters homology. With a larger amount of points, the average edge length decreases, while the distance to bisect the circle’s center, the primary way to alter homology in our class of manifolds, stays constant. Hence, increasing $n$ approaches the threshold distance in expectation, causing the correct homology of $S^1$ to appear for a larger fraction of distance parameters.

4 Conclusion

In this paper, we have demonstrated the use of persistent homology to reliably extract topological information from data. A basic hypothesis test on homological output was able to discover hollowness in data, which can be generalized to any number of dimensions. The test has the benefits of conservative rejection regions, robustness to sampling noise, and consistency against diffeomorphisms for the examples considered.

Of course, this is a first step in studying the application of persistence to manifold learning. Both fields are active in their research, and possibilities for their synthesis continue to grow. One particular area of interest for future research is zig-zag persistence [4], which calculates homology over partitions of the space data sits in, as well as intervals of distance parameters. This allows one to discover hollowness in data resembling Swiss cheese, that is, with multiple holes.

Also the homology of a sphere is but one topological signature. It will be interesting to examine if other manifolds have their own statistical interpretation.

Acknowledgments

This work was funded by a grant from the DMS-NSF VIGRE program to Stanford Statistics departments. We are thankful to Gunnar Carlsson, Persi Diaconis, Dmitriy Morozov, and Jennifer Kloke for enlightening discussions.

References