Baby Morse Theory for Statistical Inference from Point Cloud Data

CAREN MARZBAN
Applied Physics Laboratory, and Department of Statistics, University of Washington, Seattle, WA 98195
and
ULVI YURTSEVER
MathSense Analytics, 1273 Sunny Oaks Circle, Altadena, CA 91001, and Hearne Institute for Theoretical Physics, Louisiana State University, Baton Rouge, LA 70803

A methodology is proposed for inferring the topology underlying point cloud data directly from point cloud data. The approach employs basic elements of Morse theory, and is capable of producing not only a point estimate of various topological quantities (e.g., genus), but it can also assesses their sampling uncertainty in a probabilistic fashion. Several examples of point cloud data in three dimensions are utilized to demonstrate how the method yields interval estimates for the topology of the data as a 2-dimensional surface embedded in R^3.

Categories and Subject Descriptors: I.3.5 [**Computational Geometry and Object Modeling**]: Constructive solid geometry

General Terms: Algorithm

Additional Key Words and Phrases: Topology, Morse theory, persistence, inference, shape modeling, signal detection.

1. INTRODUCTION

There are many sources of high-dimensional data that are inherently structured but where the structure is difficult to conceptualize. The motivation to organize, associate, and connect multidimensional data in order to qualitatively understand its global content has recently led to the development of new tools inspired by topological methods of mathematics [Carlsson 2009; Fleishman et al. 2003; Friedman 1998; Gal and Cohen-Or 2006; Hart 1998; Ni, Garland, and Hart 2004; Niyogi, Smale, and Weinberger 2008; Pastore et al. 2006; Patena, Spagnuolo, and Falcidieno 2009; Pauly, Kobbelt, and Gross 2006; Wood et al. 2004; Zomorodian 2005]. The applications of topological data analysis methods include dimensionality reductioni [Lee and Verleysen 2007], computer vision [Pascucci et al. 2010], and shape discovery [Adan et al. 2000; Bronstein et al. 2011].In most of these applications, the data are point cloud data, i.e., the coordinates of points in some space. Such data arise naturally in LIDAR (Light Detection and Ranging) [Grejner-Brzezinska and Toth 2003], image reconstruction [Hajihashemi and El-Shenawee 2008], and in the geosciences [Wawrzyniec et al. 2007]. In addition, point-cloud data in multidimensional Euclidean space can arise from nonlinear transforms of other kinds of processes such as time series [Gilmore and Lefranc 2002].

Consider, for example, a cloud of points in 3-dimensional Euclidean space. The cloud of points may be confined mostly to the surface of a 2-dimensional sphere; or to the surfaces of multiple disconnected spheres. The number of such spheres is an example of a topological quantity, in contrast to the specific shape of the spheres (e.g. round vs. squashed) which is a geometrical quantity. Another example of a topological quantity is the number of handles; a sphere has none, but the surface of a doughnut has one. A sphere and a doughnut are topologically distinct surfaces in that one cannot be transformed to the other without cutting and gluing operations. The number of handles, known as *genus*, is important because it turns out any 2-dimensional compact surface can be constructed by gluing handles onto a sphere [Lee 2000]. Said differently, the genus is a defining characteristic of the topology of a 2-dimensional compact surface.

Whereas the human eye is capable of inferring such structures, one often requires a method for performing that task objectively. For instance, the high dimensionality of the data may not allow visualization in 3 dimensions. Even in 3 dimensions, it may be that the topological structure must be inferred in a streaming environment, where a human operator cannot visually inspect every situation one at a time. Finally, there may be situations wherein the existence of an underlying structure is not unambiguously evident even to a human expert. In such a situation, an algorithm capable

of assigning probabilities to the various topological structures can be useful for decision making [Katz and Murphy 1997].

Inferring the various disconnected components of any structure can be done via a class of statistical methods generally known as cluster analysis [Everitt 1980]. Some cluster analysis methods are also naturally capable of assigning probabilities to the different number of components/clusters. However, such methods are incapable of inferring higher-order topological structures. For instance, no clustering algorithm can identify the number of handles (i.e. genus) of a 2-dimensional surface underlying point cloud data. It is that task which is addressed in this paper. The method also produces probabilities for different genus values.

Two main methodologies of topological data analysis have been discussed so far in the literature: One is based on the idea of persistence [Bubenik et al. 2010], and the other on discretized approaches to Morse theory [Ni, Garland, and Hart 2004; Hart 1998; Patena, Spagnuolo, and Falcidieno 2009; Wood et al. 2004]. Persistence captures topological features in data by analyzing continuous structures associated to data points as a function of a varying scale parameter that measures, roughly, how coarsely the data points are assumed to sample an underlying topological manifold. While persistence relies on sophisticated constructions derived from algebraic topology, Morse theory supplies the set of tools for an alternative approach to topological data analysis [Bremer and Pascucci 2007; Cazals et al. 2003; Connolly 1996; Hart 1998; Ni, Garland, and Hart 2004; Nicolaescu 2010]. Morse theory, in its simplest form, can be thought of as a set of topological constraints which must be satisfied by a surface, if/when some function on the surface is known. For example, consider a circle (i.e., a 1-dimensional, compact surface) in 3 dimensions, oriented along the conventional z-axis. Also, consider the height function on such a circle; it is a function defined on the circle which produces the height of every point on the circle from the x-y plain. Such a function has two critical points, at the bottom and at the top of the circle, where its derivative is zero. These critical points of the height function restrict the topology of the surface over which the function is defined. In particular, they allow one to infer the genus of the underlying surface.

Many application of Morse theory are based on the ultimate desire to infer the precise shape of an underlying "object," and for this reason the theory is applied *after* a mesh has been introduced on the object [Gal and Cohen-Or 2006; Wood et al. 2004]. In the present work, Morse theory is applied directly to the point cloud data, without the need to introduce a mesh at all. This is beneficial in situation where the genus is of interest independently of finer specifications such as the shape of the object. Applying Morse theory directly to the point cloud data also provides a framework conducive to statistical inference, because a probabilistic estimate of the topology follows naturally. Specifically, in the current work, resampling [Efron and Tibshirani 1993; Good 2005] is employed to compute the empirical sampling distribution of the genus (and Betti numbers), which in turn allows for a probabilistic assessment of topology. Although not for the purpose of inferring genus, Bayesian methods have also been introduced for surface reconstruction [Diebel, Thrun, and Brunig 2006].

The contributions of this work are twofold. First, it is shown that Morse theory can be employed to infer the topology of the manifold underlying point cloud data without the introduction of a mesh, i.e., from the point cloud data itself. In the examples, which are point clouds in R^3 , the objects/manifolds are 2-dimensional surfaces and their topology is uniquely set by one integer: the genus. Second, we point out that the genus (and more generally, all algebraic topological invariants of the data) can be treated as a random variable when inferred from data. A resampling method is employed to compute the empirical sampling distribution of the genus, which in turn, conveys its sampling variability. As such, one can predict the underlying topology in a probabilistic fashion. The effect of noise on the precision of the estimates is also examined. A version of this work, but with fewer examples and without the analysis of the noise, has been presented in [Marzban and Yurtsever 2011].

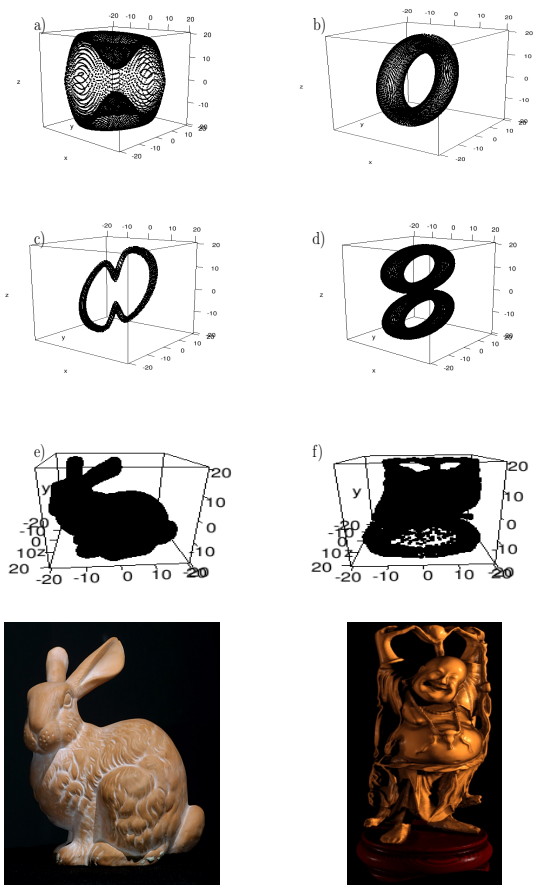

Figure 1. Six example point clouds: a) a "dimpled sphere" (genus = 0), b) a torus (genus = 1), c) a "dimpled torus" (genus=1), d) a 2-torus (genus = 2), e) the Stanford Bunny, and f) Buddha. Images of the Bunny and the Buddha are also shown to aid the visualization of their point cloud data.

2. METHOD

2.1 Generalities

To demonstrate the methodology, four simulated compact surfaces are considered, plus two often-used but more-realistic examples - the Stanford Bunny and the Happy Buddha[1] all shown in Figure 1. The choice of the simulated examples is based on the desire to

[1] http://graphics.stanford.edu/data/3Dscanrep/

have nontrivial topology, but also sufficiently simple topology to allow for a lucid presentation. Figure 1a is topologically a sphere. However, two "dimples" are introduced in order to generate more critical points for the height function, rendering the problem less trivial. Figure 1b shows the next nontrivial example, namely a torus. These two surfaces have genus 0 and 1, respectively. The next example (Figure 1c) is a genus 1 surface, but with "dimples," again for the purpose of having a more complex height function. The final simulated example (Figure 1d) is a 2-torus, i.e., a genus 2 surface. Figure 1e and 1f display the point cloud for the Bunny and the Buddha; to aid the visualization of these point cloud data, images of the two objects are included on the last row of Figure 1. In spite of its complex embedding in R^3, the topology of the Bunny is that of a sphere, i.e., genus = 0, and the genus for Buddha is 6 [Wood et al. 2004].

The particular embeddings/shapes of the simulated surfaces shown in Figure 1 are employed in the remainder of the article. Other embeddings/orientations lead to different height functions; alternatively, functions other than the height coordinate can be used to assess the topology. The particular embedding of the Bunny and the Buddha are shown Figures 4 and 5, respectively. The discussion section addresses the effect of changing the embedding for the specific purpose of obtaining more precise (less variable) estimates of the genus.

Point cloud data are simulated by adding a zero-mean random Gaussian variable to the height function of the surfaces. The variance of this variable controls the level of noise in the data. Naturally, and as shown here, small values generally lead to accurate and precise estimates of genus. Said differently, the inferred value of genus is the correct one, and the uncertainty of the estimate is small. Although larger values of the variance are associated with less precise estimates of genus, for sufficiently large values the estimates become inaccurate as well, in the sense that the most likely genus inferred from data is the wrong genus altogether. An analysis of the sensitivity of the method to noise level is sufficiently complex to be relegated to a separate article (reported later). The complexity of that analysis arises because the effect of noise level is confounded with the relative size of the various loops around the handles. For example, even with low noise levels, if one of the tori in the 2-torus is much smaller than the other, then the method is likely to imply that the underlying surface has genus one. For the present work, the standard deviation of the noise added to all of the examples, except for the Buddha, is fixed at 0.1. Loosely speaking, given that the radius of the small loop in the torus example is 4 (grid lengths), a standard deviation of 0.1 corresponds to a signal to noise ratio of about 40. For the Buddha, four different noise levels are examined in order to demonstrate the effect of noise on the sampling distribution of the genus.

Morse theory requires knowledge of the number of minima, saddle points, and maxima. There exist numerous techniques for finding critical points of a function, but in this article a relatively simple approach is adopted, again for the sake of clarity.

2.2 Specifics

Although the height function is a standard function on a surface [Bott 1980; Nicolaescu 2010], the function adopted in this article is the *area* of the surface up to some height h, denoted $S(h)$. The area function is closely related to the height function, but is more natural when dealing with data. First, the height function for data is more noisy than the area function, because the latter is inherently an integral. Second, critical points of the height function correspond to points in $S(h)$ where the derivative $S'(h)$ is discontinuous. The more robust nature, and the presence of "kinks" in the area function make it a natural choice to use for identifying the critical points in the height function.

Given that $S(h)$ is computed from data, it is a random variable. In other words, every realization of the Gaussian about the surface will lead to a different value. In order to assess the variability of $S(h)$ resampling is employed [Efron and Tibshirani 1993; Good 2005]. Specifically, 100 samples/realizations are drawn and the distribution of $S(h)$, at prespecified values of h is generated. Each distribution is summarized with a boxplot and displayed for all h values as a means of displaying the functional dependence of $S(h)$, as well as its variability, on h.

Note that each sample/realization of data gives rise to a sequence of $S(h)$ values at prespecified h values. As such, $S(h)$ can be considered a stochastic time series. Additionally, it is a monotonic, non-decreasing time series. This monotonic nature of the time series makes it difficult to identify its kinks (i.e., critical points of the height function). A more useful quantity is the first derivative of $S(h)$ with respect to h. Second derivatives are also useful, but here only the time series of the first derivatives, $S'(h)$, is examined. It is the critical points of the $S'(h)$ time series which are used in Morse theory to infer topology. The sampling variability of $S'(h)$ is again assessed via resampling, and displayed with boxplots.

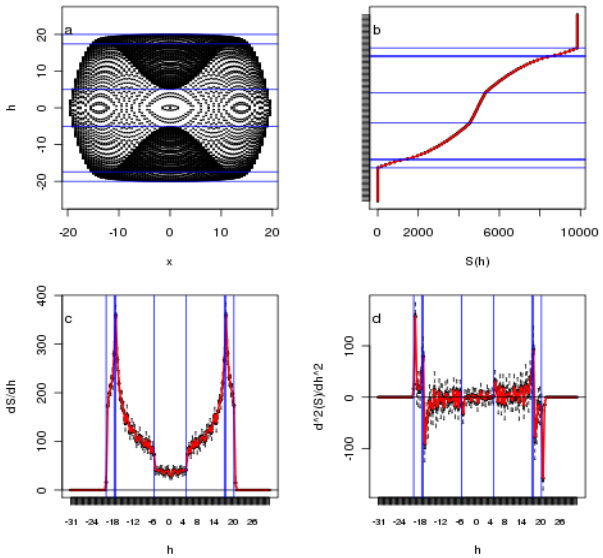

Figure 2. a) A vertical cross-section of the dimpled sphere shown in Figure 1. The blue lines mark the height of the critical points. b) The dependence of the area function $S(h)$ on the height h shown along the y axis. The blue horizontal lines mark the height of the critical points. c) The first derivative of $S(h)$ with respect to h, and the second derivative in panel d).

Figure 2 shows the above ideas for the specific example of a dimpled sphere. Figure 2a shows a vertical cross-section of the surface. Here the h values vary from the global minimum of the surface to its global maximum, in increments of 0.5. The data simulated about this surface are not shown, but boxplots summarizing the distribution of the $S(h)$ are shown in Figure 2b. Although the boxplots are relatively small, and difficult to see, their medians are connected by a red line as a visual aid. Also difficult to see are the "kinks" in the red line at the critical points, marked by the blue horizontal lines. The first derivative (Figure 2c) better shows both the kinks and the sampling variability. It is evident that some kind of a kink exists at each of the critical points of $S'(h)$ (again, marked by the

blue lines). The kinks can be generally classified into three types: an increasing step function, a cusp (i.e., $\wedge$), and a decreasing step function, respectively corresponding to minima, saddle points, and maxima. The second derivative of $S(h)$ is also shown (Figure 2d), only to illustrate that it too carries information useful for identifying critical points. However, it is not used in the present work.

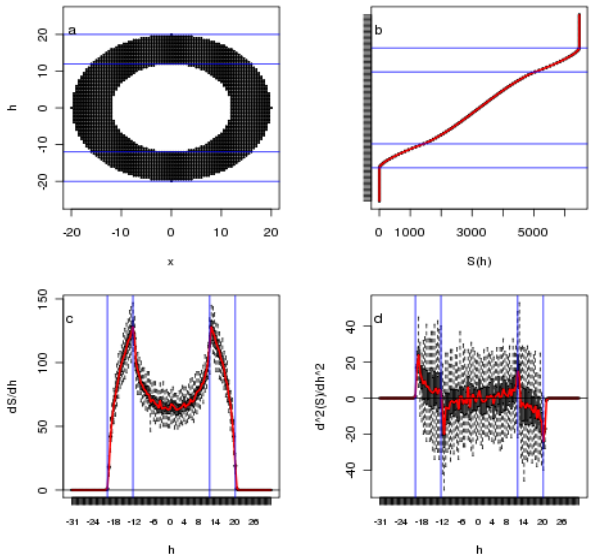

Figure 3. Same as Figure 2, but for the torus.

The analogous figures for the torus example are shown in Figure 3. Again, it can be seen that the kinks in the area function (and its derivatives) occur at the locations of the critical points of $S'(h)$, and that the shape of the kinks in the first derivative are of the same type as seen previously, namely step functions, and cusps. Similar results are found for the dimpled torus and the 2-torus (not shown). The analogous figures for the Bunny and the Buddha are shown in Figures 4 and 5.

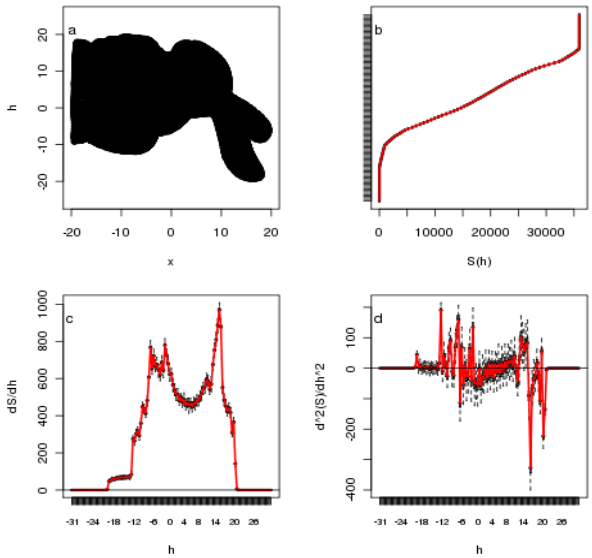

Figure 4. Same as Figure 2, but for the Bunny.

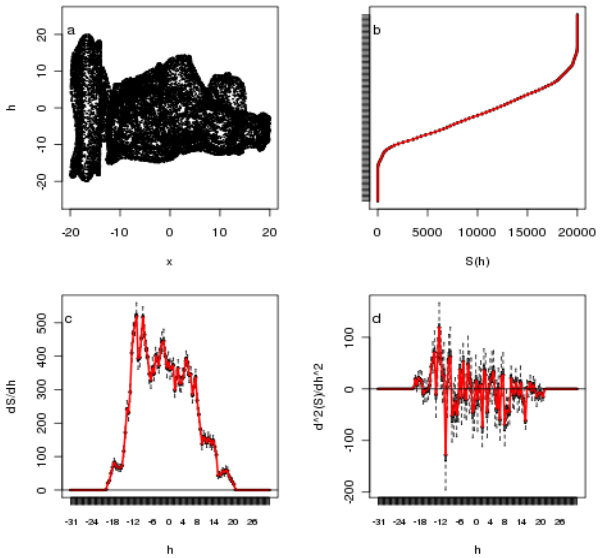

Figure 5. Same as Figure 2, but for the happy Buddha.

2.3 Finding Critical Points

Although there exist standard methods for finding critical points of a time series, most rely on some sort of time series modeling. The time series models, in turn, have numerous parameters which must be determined. Although there exist criteria (e.g., maximum likelihood) for estimating the best models, for the sake of clarity, a very simple approach is adopted here. The approach is based on template matching. Specifically, three templates are selected corresponding to the aforementioned three kinks observed in the series $S'(h)$, namely 1) an increasing step function for finding local minima in the time series, 2) a cusp function for finding the saddle points, and 3) a decreasing step function for identifying local maxima in the series.

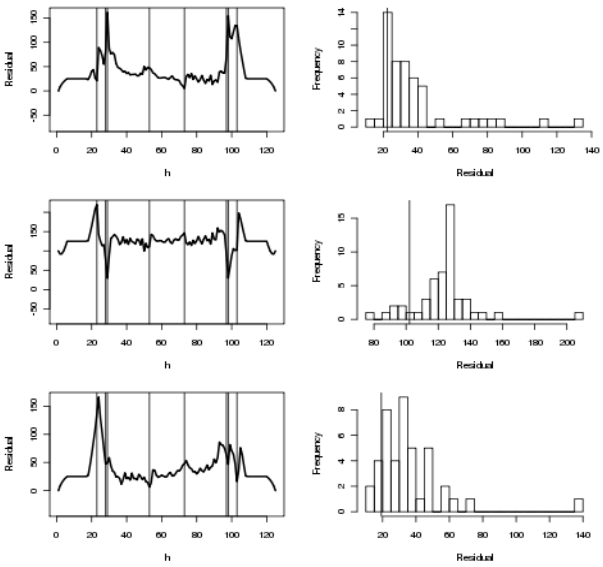

Figure 6. Left column: The time series generated by sliding three template across the time series of $S'(h)$ and computing a measure of the error/residual between the time series and each

template. Right column: The histogram of the three template errors. From top to bottom, the templates are the increasing setp function, the cusp, and the decreasing step function.

By sliding each of the templates across the time series for $S'(h)$, and computing the residuals, one obtains three additional time series. The left column in Figure 6 shows these series for one realization of data about the dimpled sphere. The vertical lines are at the h values corresponding to the critical points. Given that these time series are of residuals, near-zero values indicate a close agreement between the template and the time series of $S'(h)$. It can be seen that the residuals corresponding to the first template (top/left panel in Figure 6) approach zero only at the location of the local minima. Similarly, the residuals for the second template (middle/left panel) are near zero only at the location of the saddle points. The final panel shows the residuals for the last template, and the residuals are near zero only at the location of the local maxima. To quantify the notion of "near-zero," the histogram of the residuals is examined (right column in Figure 6). Specifically, any residual less than one standard deviation of zero is defined to be "near-zero." This 1-standard-deviation value is displayed with the vertical line on the histograms in Figure 6.

In short, sliding three templates across the time series of $S'(h)$, and examining near-zero values of the ensuing residuals correctly identifies the locations of the critical points of $S(h)$. This method for automatically identifying critical points of the height function for data can be improved upon. And as mentioned previously, there exist more sophisticated methods for identifying critical points. However, that is not the main goal of the present work. The rudimentary method outlined here is sufficient to demonstrate the main goal of the work - that Morse theory can be employed to estimate the topology directly from point cloud data, and to express the statistical uncertainty in that estimate.

3. MORSE THEORY

The material presented in this subsection is only a small portion of Morse theory, and so, has been called Baby Morse Theory [Bott 1980, 1982].

Given a surface S, the Poincare polynomial is defined as

$$P(S) = \sum_k \; b_k \; t^k \; ,$$

where $-1 \leq t \leq 1$, is a quantity with no special meaning, and b_k is the k^{th} Betti number. For a 2-dimensional surface, $k = 0, 1, 2$. Intuitively, b_0 is the number of simply-connected components of S, b_1 is the number of noncontractable loops on the surface, and b_2 is the number of noncontractable surfaces. For example, for a 2-sphere, $P(S) = 1+t^2$, and for a torus, $P(S) = 1+2t+t^2$. The $2t$ term reflects the fact that there are two noncontractable loops on a torus - one around the "hole" of the doughnut, and another going around the "handle." As another example, consider a 2-torus for which $P(S) = 1 + 4t + t^2$. It is important to point out that $P(S)$ is a topological quantity in the sense that any 2-sphere (symmetric, squashed, dimpled, or otherwise) has $P(S) = 1 + t^2$. The same is true of the other examples considered; their Poincare polynomial is independent of their embedding/shape.

Given a function f defined on a surface, the Morse polynomial is defined as

$$M(f) = \sum_{P_i} \; t^{n_i} \;\; ,$$

P_i denotes the critical points of f, and n_i is the index of f at the i^{th} critical point. The index is defined to be the number of non-decreasing directions for f. Unlike the Poincare polynomial, the Morse function is not a topological quantity. For example, consider the perfectly round 2-sphere. Then the height function has 2 critical points, with indices 0 and 2, corresponding to the South and North poles, respectively. This is so, because at the South pole there are no directions in which the height function decreases, while there are two such directions at the North pole. Then, for the height function on this sphere one has $M(f) = 1 + t^2$. By contrast, a 2-sphere with dimples in it (e.g., Figure 1a) has 6 critical points with indices 0, 1, 2, 0, 1, 2, respectively, moving up from the bottom of the figure. For this height function, $M(f) = 2 + 2t + 2t^2$. As another example, for the height function on the torus in Figure 1b, one has $M(f) = 1 + 2t + t^2$.

Central to Morse theory are the so-called Morse inequalities [Bott 1980; Nicolaescu 2010]. They are expressed in two forms - "weak" and "strong:"

$$M(f) \geq P(S) \quad , \quad M(f) - P(S) = (1+t)Q(t), \qquad (1)$$

where $Q(t)$ is any polynomial in t with non-negative coefficients.

In the above examples, note that for some functions one has $M(f) = P(S)$. Such functions are called "perfect." Intuitively, such a function tightly "hugs" the surface. As such, the coefficients in the corresponding Morse function are equal to the Betti numbers. As a result, knowledge of a perfect function is tantamount to precise knowledge of the topology (technically, homology) of the underlying surface. For all non-perfect functions, the Morse inequalities provide only an upper bound on the Betti numbers, and do not uniquely identify the topology.

The search for perfect functions is aided by the Lacunary principle [Bott 1980]: If the product of all consecutive coefficients in $M(f)$ is zero, then f is perfect. Another useful corollary of the strong form of the inequalities follows upon considering $t = -1$:

$$\sum_{P_i} (-1)^{n_i} = \sum_k \; b_k \; (-1)^k. \qquad (2)$$

This places a constraint on the allowed number of minima, saddle points, and maxima:

$$n_{min} - n_{saddle} + n_{max} = b_0 - b_1 + b_2. \qquad (3)$$

And since in this article only surfaces with $b_0 = b_2 = 1$ are considered, then

$$b_1 = 2 - n_{min} + n_{saddle} - n_{max} \; . \qquad (4)$$

Finally, given that any 2-dimensional surface can be constructed by gluing tori to a sphere, it follows that b_1 must be even (including zero). The genus of a compact surface is then found to be

$$\text{genus} = b_1/2 \; . \qquad (5)$$

For a sphere, a torus, and a 2-torus (e.g. in Figure 1), the genus is 0, 1, and 2, respectively. Intuitively, the genus counts the number of "holes" or "handles" in a compact surface.

4. RESULTS

Armed with a method to find the number of minima, saddle points, and maxima of the height function (section 2.3), one can then examine the distribution of each. The top panel in Figure 7a shows the boxplots summarizing the three distributions for the dimpled sphere example. Recall that for this example, the correct number of minima, saddle points, and maxima is 2 for each. The median of the three boxplots is found to be precisely at 2. The 1st and 3rd quartiles of the distribution (i.e., the bottom and top sides of the boxes

in the boxplots) suggest an uncertainty of about ± 1 for each of the numbers. In other words, the number of minima, saddle points, and maxima generally varies within 1 of the correct value (i.e., 2).

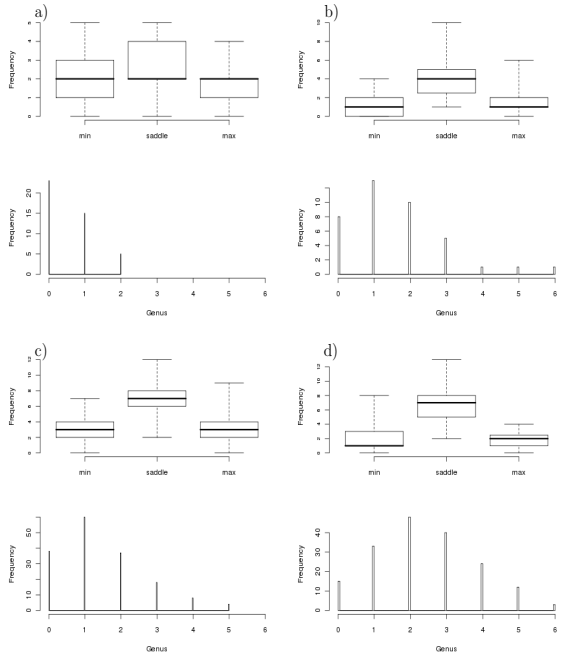

Figure 7. The boxplot summary of the distribution of the number of minima, saddle points, and maxima, and the histogram of the estimated genus. The 4 panels pertain to the four examples: a) dimpled sphere, b) torus, c) dimpled torus, and d) 2-torus.

However, not all of the values in that range are allowed. Eq. (4) constraints the three numbers, because the first Betti number must be even. This constraint reduces the uncertainty even further. Meanwhile, the main interest is in the value of the genus, which can be computed from Eq. (5). The histogram of the genus is shown immediately below the comparative boxplots in Figure 7. Interpreting this histogram probabilistically, it can be seen that the most likely value of the genus is zero. And, of course, that is the correct value. Moreover, values of estimated genus as large as 2 are possible, but less likely.

The remaining panels in Figure 7 show the analogous figures for the torus (Figure 7b), dimpled torus (Figure 7c), and the 2-torus (Figure 7d). The correct number of minima, saddle points, and maxima for the torus is (1,2,1). The analogous numbers for the dimpled torus and the 2-torus are (3,6,3), and (1,4,1), respectively. The comparative boxplots in Figure 7, are all in agreement with these numbers. It is worth noting that the spread of the boxplots generally increases with the complexity of the underlying surface.

The distribution of the estimated genus for all four simulated examples is also consistent with the correct values (Figure 7). The most likely genus for the torus, dimpled torus, and 2-torus are 1, 1, and 2, respectively - the correct values. As with the number of critical points, the uncertainty in the estimated genus increases with the complexity of the surface. Whereas the genus for the dimpled sphere varies between 0 and 2, the range for the 2-torus is 0 to 6.

As for the Bunny and Buddha examples, the true number of minima, maxima, and saddle points is not known, and so, a comparison of their distributions with the true values is not possible. The distribution of estimated genus for the two examples is shown in Figure 8. The most likely value of genus for the Bunny is zero, i.e., the correct value. However, that for the Buddha is 5, one less than the correct value. The reason for this error is complex, and points at an important issue discussed next.

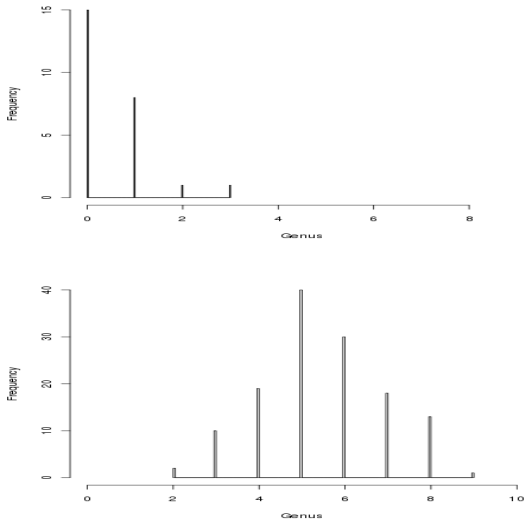

Figure 8. The empirical sampling distribution of genus for the bunny (top) and for Happy Buddha (bottom). The most likely genus for these examples are 0 and 5 (not 6), respectively.

Recall that all of the surfaces examined thus far are contaminated by a zero-mean gaussian with $\sigma = 0.1$. In order to examine the sampling distribution of the genus for different noise-levels, the Buddha example is analyzed with four different values of σ, namely, 0.01, 0.05, 0.1, and 0.5. The results are shown in Figure 9. First, note that the width of the sampling distribution of genus increases with increasing noise level. This is a desirable feature of the method, because one's confidence in the inferred estimate ought to diminish with increasing noise.

Focusing on the modes of the distributions, for the lowest noise level ($\sigma = 0.01$), the most likely value of genus is 7, one higher than the correct value. For $\sigma = 0.05$ the most likely genus is 6, i.e., the correct value. As σ is increased to 0.1 and 0.5, the most likley value of genus falls to 5 and 4, respectively. The reduction of genus with increasing noise is not surprising; as the noise level increases, some of the smaller "holes" in the surface disappear, thereby reducing the genus. The explanation of why the most likely genus is larger than the correct value, for the lower noise levels, is related to the choice of the parameters of the templates used for identifying critical points; this issue is discussed in the next section.

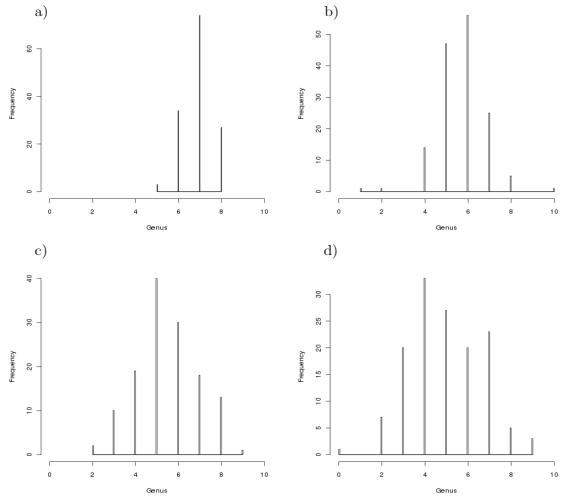

Figure 9. The empirical sampling distribution of genus for the Happy Buddha for different levels of noise; $\sigma = 0.01, 0.05, 0.1, 0.5$**, for a-d, respectively.**

5. SUMMARY AND DISCUSSION

The Morse inequalities are reviewed. It is shown that when specialized to the case of a 2-dimensional surface embedded in 3 dimensions, they place severe constraints on the topology of the surface. Several examples are employed to show that all of the quantities appearing in the Morse inequalities can be estimated directly from point cloud data, thereby providing a statistical/probabilistic view of the topology of the surface underlying the data. Empirical sampling distributions are produced for the various topological entities, all of which can then lead to traditional confidence intervals or hypothesis tests of the topological parameter of interest.

In a statistical setting, i.e., where an object has been sampled with some density, the notion of the "correct" value of its genus is ambiguous, at best, because the estimated genus depends on the scale at which the data is examined. The aforementioned "correct" values are all based on a visual inspection of the true objects, not the point cloud generated from them. Any attempt to extract genus from point cloud data, by any method, must take into account some notion of the scale of the features of interest. Methods based on persistence [Bubenik et al. 2010] are no exception, because they too involve a parameter (ϵ) which effectively controls the scale of interest. In inferring genus, one must specify the typical size of the homology loops, for example. Without some specification of the scale of interest, the inferred genus will depend on how "closely" the point cloud is examined. In the proposed method, there are several parameters that control the scale of interest. All of them appear in inferring the critical points of a function. For example, the step and cusp templates discussed in section 2.3 all have parameters that define how quickly $S'(h)$ increase or decreases, or how sharp the cusp is expected to be. Furthermore, even after specifying these quantities, one must still specify which steps or cusps in the time series of $S'(h)$ are of interest; in the present work, this was implemented by adopting a criterion which keeps only steps and cusps with errors less than 1 standard deviation of zero (right column in Figure 6). Relaxing or tightening this criterion will lead to a different set of critical points, and a different distribution for genus. In the simulated example, the typical size of the features is relatively unambiguous, and so, the proposed method yields the correct genus. But for the more complex examples such as the Buddha, the method may predict the "wrong" genus (e.g., Figure 8, bottom panel) if the scale of interest has not been specified correctly. It is also worth pointing out that the choice of the scale and the noise level are in fact confounded. As such, the spread of the sampling distribution reflects uncertainty due to both noise and misspecification of scale. The confounding nature of scale and noise are currently under investigation.

Although not shown here, we have found that the spread of the sampling distribution (i.e., uncertainty) of the genus generally depends on the orientation of the surface. This is expected, because the height function depends on the orientation. So, it is possible to orient the surface in a way that would allow for more precise estimates of the critical points. In other words, it is possible to add another step to the proposed method, wherein the variance of the distribution of genus is minimized across different orientations of the point cloud. Such a rotation can also be used to identify a perfect height function, in which case the Betti numbers can be computed exactly, as opposed to being only bounded at the top. This idea will be examined in a later publication.

It is also important to note that while we focus in this paper on topology estimation, the critical points and indices of naturally defined functions over the data (such as height functions) give more than topological information: For example, for the dimpled torus (Figure 1c), the topology is identical to that of a standard torus (Figure 1b), but the critical points of $S'(h)$ do give information about the dimples, i.e., a geometric rather than topological feature of the data set. The extra geometric information extracted from data via Morse theory might be useful for some applications such as molecular shape determination [Cazals et al. 2003].

In the examples considered here the goal is to identify the genus of a 2-dimensional compact surface underlying 3-dimensional point cloud data. Several generalization are possible. The dimensionality of the embedding space, or of the "surface" (embodying the underlying structure), can both be generalized. Of course, a single number like genus will no longer suffice to define the topology uniquely, but the set of Betti numbers does. In other words, if the manifold of interest underlying the data has dimension larger than 2, then more parameters need to be estimated. From a statistical point of view, the consequence of this increase in the number of parameters is that more data will be required to estimate the parameters with precision.

The general formulation of Morse theory does not require the underlying manifold to be compact. There are also extensions of Morse theory that allow for degenerate critical points, as well as extensions to manifolds with boundary, and to Morse functions that take values in more general spaces than $\mathbf{R}$ (e.g., circle-valued Morse theory where Morse functions are S^1-valued) [Pajitnov 2006]. The application of these more powerful topological tools to data analysis is a fruitful frontier for exploration.

Acknowledgements

Marina Meilă is acknowledged for valuable discussion. support for this project was provided by the National Geospatial-Intelligence Agency, award number HM1582-06-1-2035.

REFERENCES

Adan, A., Cerrada, C., Feliu, V. 2000. Modeling Wave Set: Definition and Application of a New Topological Organization for 3D Object Modeling, *Computer Vision and Image Understanding 79*, 281-307.

Bott, R. 1980. Morse theoretic aspects of Yang-Mills Theory, in *Recent Developments in Gauge Theories*, Eds. G. 'tHooft, et al., Plenum Publishing, pp. 46-67.

Bott, R. 1982. Lectures on Morse Theory, Old and New. *Bull. Amer. Math. Soc. (N.S.), 7*, 331-358.

Bremer, P-T. , Pascucci, V. 2007. A Practical Approach to Two-Dimensional Scalar Topology, in *Topology-based Methods in Visualization*, H. Hauser, H. Hagen, H. Theisel (Eds), Springer-Verlag, Berlin. ISBN-13 978-3-540-70822-3.

Bronstein, A. M., Bronstein, M. M., Guibas, L. J., Ovsjanikov, M. 2011. Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Trans. Graph. 30*, 1, 1-22.

Bubenik, P., Carlson, G., Kim, P. T., Luo, Z-M. 2010. Statistical Topology Via Morse Theory Persistence and Nonparametric Estimation. *Algebraic Methods in Statistics and Probability II*, 516-575.

Carlsson, G. 2009. Topology and Data. *Bull. (New Series) Amer. Math. Soc., 46*, 255308.

Cazals, F., Chazal, F., Lewiner, T. 2003. Molecular Shape Analysis Based upon the Morse-Smale Complex and the Connolly Function. *Proc. 19th ACM Symp. Computational Geometry (SoCG)*, 351-360.

Connolly, M. 1996. Molecular Surfaces: A Review. *Network Science*. Available at http://www.netsci.org/Science/Compchem/feature14.html.

Diebel, J. R., Thrun, S., Brunig, M. 2006. A Bayesian method for probable surface reconstruction and decimation. *ACM Trans. Graph. 25*, 1, 39-59.

Efron, B., Tibshirani, R. J. 1993. *An introduction to the bootstrap*. Chapman & Hall, London.

Everitt, B. S. 1980. *Cluster Analysis*. Second Edition, Heinemann Educational Books. London.

Friedman, J. 1998. Computing Betti Numbers via the Combinatorial Laplacian. *Algorithmica 21(4)*, 331-346.

Fleishman, S., Cohen-Or, D., Alexa, M., Silva, C. T. 2003. Progressive point set surfaces. *ACM Trans. Graph. 22*, 4, 997-1011.

Gal, R., Cohen-Or, D. 2006. Salient geometric features for partial shape matching and similarity. *ACM Trans. Graph. 25*, 1, 130-150.

Gilmore, R., Lefranc, M. 2002. *The Topology of Chaos*, Wiley-Interscience, New York. ISBN 0-47 1-40816-6

Good, P. I. 2005. *Introduction to Statistics Through Resampling Methods and R/S-PLUS*, Wiley. ISBN 0-471-71575-1

Grejner-Brzezinska, D., Toth, C. 2003. Deriving Vehicle Topology and Attribute Information Over Transportation Corridors From LIDAR Data. *Proceedings of the 59th Annual Meeting of The Institute of Navigation and CIGTF 22nd Guidance Test Symposium*, June 23 - 25, Albuquerque, NM, 404-410.

Hajihashemi, M. R., El-Shenawee, M. 2008. Shape Reconstruction Using the Level Set Method for Microwave Applications. *Antennas and Wireless Propagation Letters 7*, 92-96.

Hart, J. C. 1998. Morse theory for implicit surface modeling. In *Mathematical Visualization*, Springer-Verlag, 257268.

Katz, R. W., Murphy, A. H. 1997. *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, Cambridge.

Lee, J., M. 2000. *Introduction to Topological Manifolds*. Springer-Verlag, New York. ISBN 0-387-98759-2

Lee, J. A., Verleysen, M. 2007. Nonlinear dimensionality reduction. *Information Science and Statistics*, Springer, 308 pp.

Marzban, C., and Yurtsever, U. 2011. Baby Morse theory in data analysis. *Workshop on Knowledge Discovery, Modeling and Simulation (KDMS)*, held in conjunction with the *17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego, CA., August 21-24.

Ni, X., Garland, M., Hart, J. C. 2004. Fair Morse functions for extracting the topological structure of a surface mesh. *ACM Trans. Graph. (Proceedings of ACM SIGGRAPH 2004) 23*, 3, 613-622.

Nicolaescu, L. I. 2010. *An Invitation to Morse Theory*. Springer Monograph, XIV. ISBN 978-0-387-49509-5. 242 pp.

Niyogi, P., Smale, S., Weinberger, S. 2008. Finding the homology of submanifolds with high confidence from random samples. *Combinatorial and Discrete Geometry 39*, 419-441.

Pajitnov, A. V. 2006. Circle-valued Morse Theory. *De Gruyter Studies in Mathematics 32*. ISBN 978-3-11-015807-6.

Pascucci, V. Tricoche, X., Hagen, H., Tierny, J. 2010. *Topological Methods in Data Analysis and Visualization; Theory, Algorithms, and Applications.* Springer, 260 pp.

Pastore, J., Bouchet, A., Moler, E., Ballarin, v. 2006. Topological Concepts applied to Digital Image Processing. *Journal of Computer Science & Technology 6*, 80-84.

Patena, G., Spagnuolo, M., Falcidieno, B. 2009. Topology- and error-driven extension of scalar functions from surfaces to volumes. *ACM Trans. Graph 29*, 1, 1-20.

Pauly, M., Kobbelt, L. P., Gross, M. 2006. Point-based multiscale surface representation. *ACM Trans. Graph. 25*, 2, 177-193.

Wawrzyniec, T. F., McFadden, L. D., Ellwein, A., Meyer, G., Scuderi, L., McAuliffe, J., Fawcett, P. 2007. Chronotopographic analysis directly from point-cloud data: A method for detecting small, seasonal hillslope change, Black Mesa Escarpment, NE Arizona. *Geosphere 3*, 550-567. DOI: 10.1130/GES00110.1

Wood, Z., Hoppe, H., Desbrun, M., Schroder, P. 2004. An out-of-core algorithm for isosurface Topology Simplification. *ACM Trans. on Graph. 23*, 2, 190-208 .

Zomorodian, A. 2005. Topology for Computing. *Cambridge Monographs on Applied and Computational Mathematics (No. 16)*.