Detecting Topological Structure

Shmuel Weinberger
University of Chicago

A number of researchers have been applying geometric techniques in the study of large data sets with the goal of obtaining more topological and qualitative information. In this talk, I will explain some of the issues involved in trying to discern geometric information from random samples, especially in the presence of noise, and give some situations where despite noise, it is possible to discover some underlying geometric structure. Such theoretical guarantees help provide tools for measuring the statistical significance of results obtained by this methodology.

TELL US, IN LAYMAN'S TERMS, WHAT YOUR BREAKTHROUGH MEANS.
CERTAINLY. $K - \frac{4n^3}{7}\sqrt{P_{..4}} \cdot \frac{\Sigma L}{5T}$
$K - \frac{4n^3}{7}\sqrt{P_{..4}} \cdot \frac{\Sigma L}{5T}$
S. Harris

Outline

1. Motivation
2. Applications
3. Theorems
4. Noise

Basic Motivating Question

Question

Is there a geometry underlying a data set, and if so, how can we detect it and how can we use it?

Basic Motivating Question

Question

Is there a geometry underlying a data set, and if so, how can we detect it and how can we use it?

Remark

Topology should have relevance because it is relatively insensitive to “small perturbations” both in terms of

- the qualitative nature of its equivalence relations and
- a set of adjectives that are crude enough to be quickly identified.

loading

Basic Motivating Question

Question

Is there a geometry underlying a data set, and if so, how can we detect it and how can we use it?

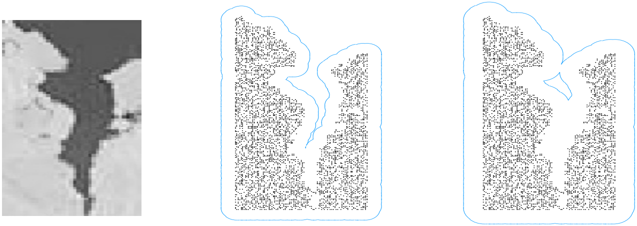
Remark

Topology should have relevance because it is relatively insensitive to “small perturbations” both in terms of

- the qualitative nature of its equivalence relations and
- a set of adjectives that are crude enough to be quickly identified.

loading

Application: Pattern recognition



V. Robins, J. Abernethy, N. Rooney, Elizabeth Bradley. Topology and intelligent data analysis. *Intelligent Data Analysis.* Volume 8, 2004. 505–515.

Here one has a universe of possible types and seeks useful invariants for distinguishing objects.

Example: Dimension.

Questions

- How many degrees of freedom are there?
- Are there any laws to be discovered?

Example: Dimension

A pair of points on the circle is equivalent to a point on the Möbius band — 2 degrees of freedom.

loading

Example: Dimension

An asymmetric rigid body on the surface of the Earth is equivalent to a point in $\mathbb{R}P^3$ — 3 degrees of freedom.

loading

Why $\mathbb{R}P^3$? Conjugating $x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}$ by a unit quaternion

$$q = \cos\theta + u\,\mathbf{i} + v\,\mathbf{j} + w\,\mathbf{k}$$

rotates (x, y, z) by angle 2θ around the axis (u, v, w).
Note that q and $-q$ induce the same rotation.

Example: Dimension

Problem

What can you do in the absence of a model for the data?

Example: Dimension

Problem

What can you do in the absence of a model for the data?

And what can you do if the data has some noise?

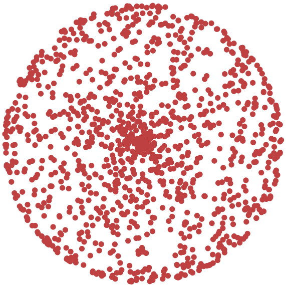
Example: Dimension

Problem

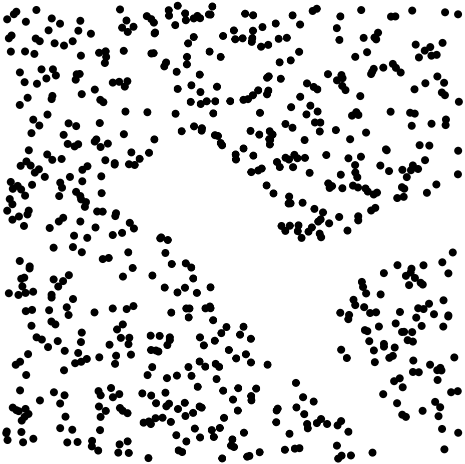
What can you do in the absence of a model for the data?

And what can you do if the data has some noise?

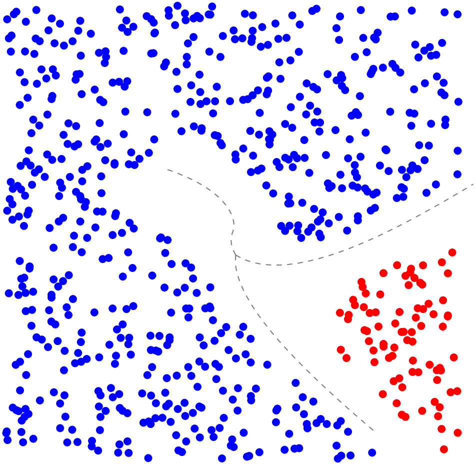
Is this the surface of a ball (a 2-sphere) or noise around a point?



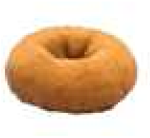
Example: Clustering



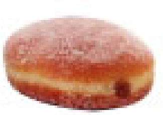
Example: Clustering



Topological type



versus



Almost all of the previous examples are *topological* properties, or are illuminated by topological invariants.

Topological invariants

Clustering is the theory of H_0, e.g., *how many components?*

Topological invariants

Clustering is the theory of H_0, e.g., *how many components?*

Dimension is closely related to the n for which all H_k vanish for $k > n$ (or $k \geq n$ for proper subsets).

Topological invariants

Clustering is the theory of H_0, e.g., *how many components?*

Dimension is closely related to the n for which all H_k vanish for $k > n$ (or $k \geq n$ for proper subsets).

Gives information about the entropy of dynamical systems on such a space.

Topological invariants

Clustering is the theory of H_0, e.g., *how many components?*

Dimension is closely related to the n for which all H_k vanish for $k > n$ (or $k \geq n$ for proper subsets).

Gives information about the entropy of dynamical systems on such a space.

Can be applied to the problem of coverage by a sensor network.

Topological invariants

Clustering is the theory of H_0, e.g., *how many components?*

Dimension is closely related to the n for which all H_k vanish for $k > n$ (or $k \geq n$ for proper subsets).

Gives information about the entropy of dynamical systems on such a space.

Can be applied to the problem of coverage by a sensor network.

Determines the topological type for two dimensional surfaces.

Justification for sampling

Theorem (Niyogi-Smale-W)

- $M \subset \mathbb{R}^N$, *compact, with condition number* τ.
- $\bar{x} = \{x_1, \ldots, x_n\}$ *uniform measure on* M.
- ϵ *small enough,* n *big enough,*
- $U =$ *the* ϵ*-neighborhood of* $\bar{x}$,

Then $H_\star(U) = H_\star(M)$ *with probability* $> 1 - \delta$.

Justification for sampling

Remark

Precise formulae connecting the dimension of the Euclidean space, the diameter of M, τ, ϵ, and δ and an algorithm for computing $H^*(U)$ are all in the original paper

> *Niyogi, Smale, Weinberger, Finding the homology of submanifolds with high confidence from random samples. Discrete Comput. Geom. 39 (2008), no. 1–3, 419–441.*

and need not bother us here.

Recovering homotopy type of a manifold from samples

Definition (Condition number.)

If M is a smooth manifold in Euclidean space, then the **condition number** $1/\tau$ of M is given by

$$\tau := \sup\{t : \text{Normal exp. on the } \epsilon\text{-normal disk bundle is 1–1}\}$$

This incorporates local curvature conditions ($\approx$ largest principal curvature) and global information about how close different coordinate charts get.

Recovering homotopy type of a manifold from samples

Definition (Condition number.)

If M is a smooth manifold in Euclidean space, then the **condition number** $1/\tau$ of M is given by

$$\tau := \sup\{t : \text{Normal exp. on the } \epsilon\text{-normal disk bundle is 1–1}\}$$

This incorporates local curvature conditions ($\approx$ largest principal curvature) and global information about how close different coordinate charts get.

a thin bagel

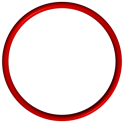
Recovering homotopy type of a manifold from samples

Definition (Condition number.)

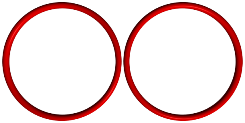
If M is a smooth manifold in Euclidean space, then the **condition number** $1/\tau$ of M is given by

$$\tau := \sup\{t : \text{Normal exp. on the } \epsilon\text{-normal disk bundle is 1–1}\}$$

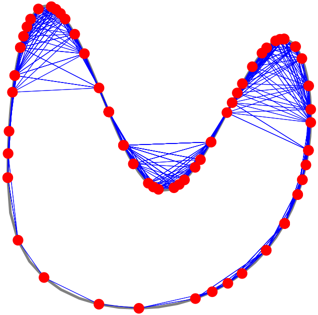
This incorporates local curvature conditions ($\approx$ largest principal curvature) and global information about how close different coordinate charts get.



versus



Sampling



Sampling can get the **homology** correct, though the model is built using high-dimensional simplices.

Remarks

- The Euclidean nature of the data set is sometimes artificial. Some parts are simplified if one uses metrics intrinsic to the underlying space—but other places we use Euclidean geometry. The latter also begs the question of estimation of intrinsic distances in terms of distances derived from the samples.

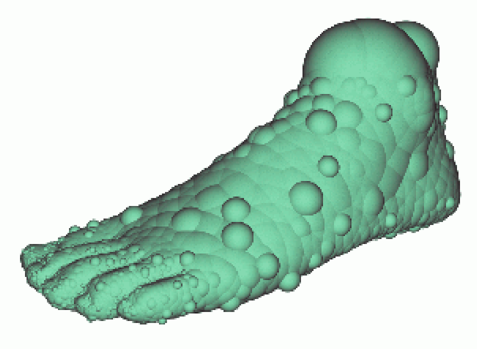
Remarks

- The Euclidean nature of the data set is sometimes artificial. Some parts are simplified if one uses metrics intrinsic to the underlying space—but other places we use Euclidean geometry. The latter also begs the question of estimation of intrinsic distances in terms of distances derived from the samples.

- This is a first stab. Later work has greatly relaxed the hypothesis of smooth manifold to allow many stratified spaces. See Cohen-Steiner, Edelsbruner and Harer and also Chazal, Cohen-Steiner, and Lieutier for such developments.

Remarks

- In the case of hypersurfaces, this method gives a topological picture of the submanifold, not just of its homotopy type.



In general such a theory was developed by Amenta in 3-dimensions, and Cheng-Dey-Ramos, and Boissonat-Guibas-Oudot.

Remarks

- This algorithm require quadratic programming, and also a hypothesis of the relevant scale.
  Edelsbrunner-Letscher-Zomorodian have an alternative called "persistent homology" that finesses this.

Remarks

- This algorithm require quadratic programming, and also a hypothesis of the relevant scale. Edelsbrunner-Letscher-Zomorodian have an alternative called “persistent homology” that finesses this.

- Using adaptive methods, it is possible to do a lot better in practice. One should not sample uniformly—if possible one should sample more sparsely areas that have less topology. (This is implicit in “Amenta’s foot” on the previous slide.)

Remarks

- This algorithm require quadratic programming, and also a hypothesis of the relevant scale.
  Edelsbrunner-Letscher-Zomorodian have an alternative called "persistent homology" that finesses this.

- Using adaptive methods, it is possible to do a lot better in practice. One should not sample uniformly—if possible one should sample more sparsely areas that have less topology. (This is implicit in "Amenta's foot" on the previous slide.)

- Adaptive methods have also been applied to homology of nodal sets by Mischaikow and his collaborators.

Remarks

- Note that all estimates are in terms of the submanifold, none in terms of the ambient Euclidean space (aside from a tacit upper bound on dimension—which requires an additional discussion—not today).

Remarks

- Note that all estimates are in terms of the submanifold, none in terms of the ambient Euclidean space (aside from a tacit upper bound on dimension—which requires an additional discussion—not today).
- It is reasonable to use the objects defined here as proxies for the homology and homotopy type of data sets even if they are not derived from a manifold. Of course, what the meaning of this homology is is then of some interest.

Remarks

- Note that all estimates are in terms of the submanifold, none in terms of the ambient Euclidean space (aside from a tacit upper bound on dimension—which requires an additional discussion—not today).
- It is reasonable to use the objects defined here as proxies for the homology and homotopy type of data sets even if they are not derived from a manifold. Of course, what the meaning of this homology is is then of some interest.
- All of this is under the assumption that our data is noise-free. We must now work to repair this. This also leads to a weakening of the assumption that the points are chosen uniformly from the submanifold.

Implementations

1. - PLEX available at Stanford `http://comptop.stanford.edu/programs/plex/`
   - CHomP available at Rutgers `http://chomp.rutgers.edu/`
2. What happens in practice?
   Joint work with Y. Baryshnikov.

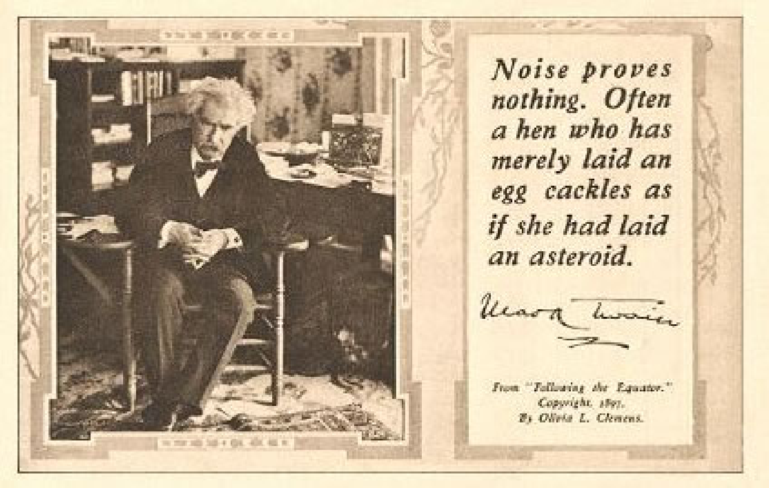
Three stages:

Dust	very little information available
Percolation	
Endgame	Getting things “right”—marked by abrupt phase transitions (following overshoot)

In practice. . .

loading

Noise



Noise proves nothing. Often a hen who has merely laid an egg cackles as if she had laid an asteroid.

Mark Twain

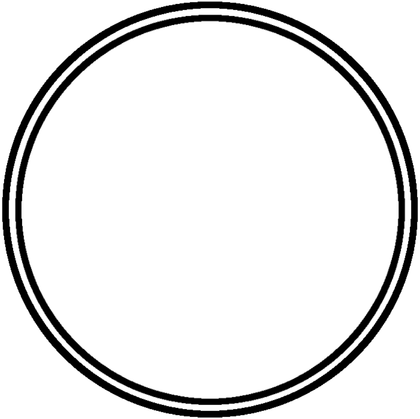
From "Following the Equator."
Copyright, 1897,
By Olivia L. Clemens.

Noise

- Problems when there is too much noise.
- Hot spots and oversampling.

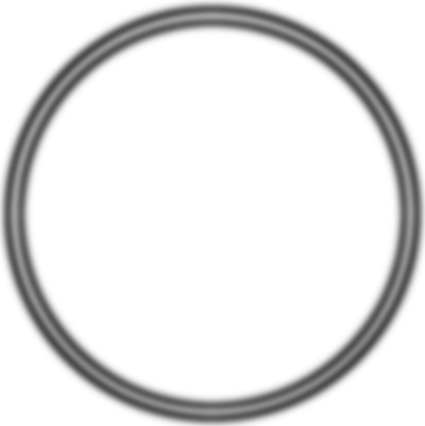
When there is too much noise. . .

Noise causes **blurring** when it is too large.



When there is too much noise. . .

Noise causes **blurring** when it is too large.



When there is too much noise. . .

Noise causes **blurring** when it is too large.

When there is too much noise. . .

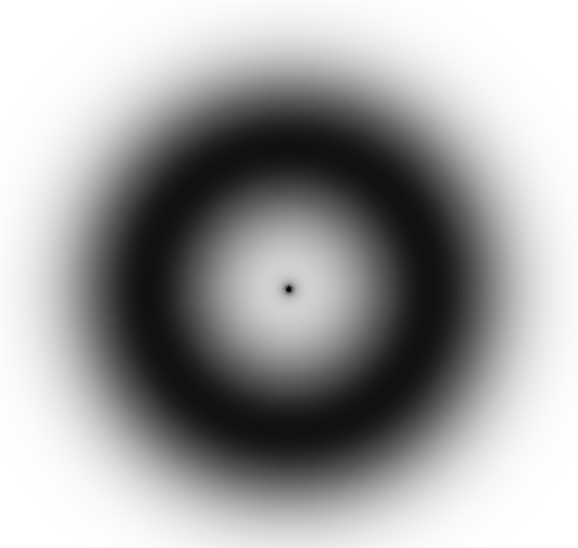
Noise causes **blurring** when it is too large.

When there is too much noise. . .

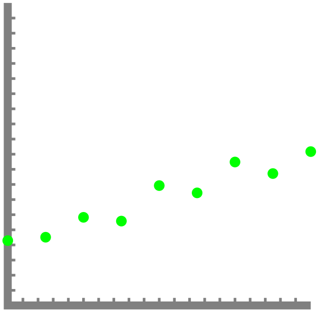
Noise causes **blurring** when it is too large.

This raises the issue of **scale**.

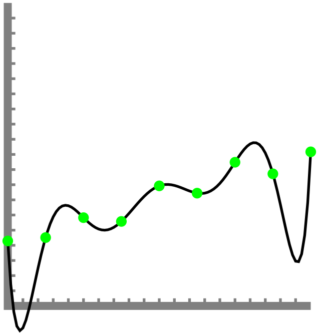
Noise and “hot spots”



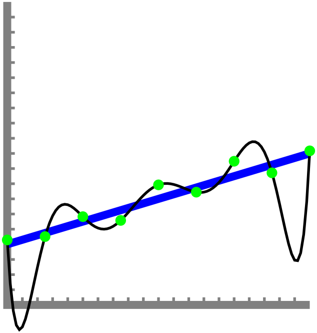
Example: Classical problem of overfitting



Example: Classical problem of overfitting



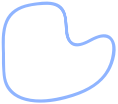
Example: Classical problem of overfitting



You must clean the data.

Slogan

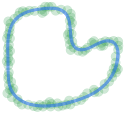
"A little noise always kills in the long run."



You must clean the data.

Slogan

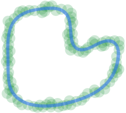
"A little noise always kills in the long run."



You must clean the data.

Slogan

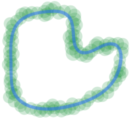
"A little noise always kills in the long run."



You must clean the data.

Slogan

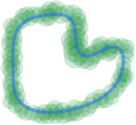
"A little noise always kills in the long run."



You must clean the data.

Slogan

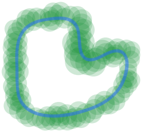
“A little noise always kills in the long run.”



You must clean the data.

Slogan

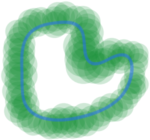
"A little noise always kills in the long run."



You must clean the data.

Slogan

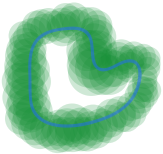
“A little noise always kills in the long run.”



You must clean the data.

Slogan

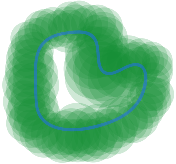
“A little noise always kills in the long run.”



You must clean the data.

Slogan

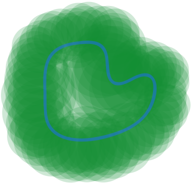
"A little noise always kills in the long run."



You must clean the data.

Slogan

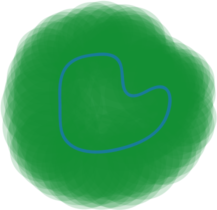
"A little noise always kills in the long run."



You must clean the data.

Slogan

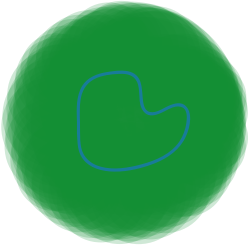
“A little noise always kills in the long run.”



You must clean the data.

Slogan

"A little noise always kills in the long run."



You must clean the data.

Slogan

"A little noise always kills in the long run."

Remark

One must clean the data.

Abstract theorem.

Theorem (Niyogi-Smale-W)

Let M be a manifold with a given τ.
Assume that for some $\epsilon < \tau/2$,
we have a measure μ satisfying:

- α *homogeneity:* $\mu(B_\epsilon(p)) > \alpha \cdot \mu(B_\epsilon(q))$ *for every* $p \in M$ *and all* q.
- β *anti-homogeneity:* $\mu(B_\epsilon(q)) < \beta \cdot \mu(B_\epsilon(p))$ *for every* $p \in M$ *and for* q *outside a* 2ϵ*-neighborhood of* M.

Thus, it is possible to clean a large enough data set and to compute the homology of M using a suitable nerve homology.

Precise theorem in the presence of Gaussian noise.

Theorem

$M^d \subset \mathbb{R}^D$. As long as the variance σ^2 satisfies

$$\sigma\sqrt{8(D-d)} < c\frac{\sqrt{9}-\sqrt{8}}{9}\tau \text{ for any } c < 1,$$

then $H_\star(M)$ can be recovered from random samples.

Remark

If the codimension is high enough,

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

for constants $A, K > 0$, then the sample complexity is independent of D.

Review on noise.

- α–β homogeneity $\Rightarrow$ can clean the data

- Gaussian noise $\Rightarrow$ sample complexity does not grow with ambient dimension

Conclusions

- If the data is not too noisy and you have enough of it, then it is possible to infer the geometric structure, e.g. find clusters, discover dimension, topological type and so on.

Conclusions

- If the data is not too noisy and you have enough of it, then it is possible to infer the geometric structure, e.g. find clusters, discover dimension, topological type and so on.
- We can infer different parts of the geometry at different rates in the noiseless case, and are working on techniques for denoising these tools.

Conclusions

- If the data is not too noisy and you have enough of it, then it is possible to infer the geometric structure, e.g. find clusters, discover dimension, topological type and so on.
- We can infer different parts of the geometry at different rates in the noiseless case, and are working on techniques for denoising these tools.
- There is a complementary subject of finding lower bounds on sample and computational complexity of these problems. Indeed these problems grow badly with the dimension of the underlying space.

Conclusions

- If the data is not too noisy and you have enough of it, then it is possible to infer the geometric structure, e.g. find clusters, discover dimension, topological type and so on.
- We can infer different parts of the geometry at different rates in the noiseless case, and are working on techniques for denoising these tools.
- There is a complementary subject of finding lower bounds on sample and computational complexity of these problems. Indeed these problems grow badly with the dimension of the underlying space.
- The good news is that "low dimensional features" can still be discovered relatively early. And more good news: The sample complexity depends on the intrinsic dimension of the space, not on the dimension of the space it is embedded in.