On the Shape of Data

Caren Marzban[1], Ulvi Yurtsever[2]

[1] Applied Physics Lab, and Dept of Statistics, University of Washington, Seattle, WA 98195
[2] MathSense Analytics, 1273 Sunny Oaks Circle, Altadena, CA, 91001

Abstract

One of the active areas in machine learning deals with the problem of determining the "shape" of data. In its simplest setting, one may be concerned with the geometric structure of spatial coordinates of data in 3-dimensional space. More abstractly, the variables in any multivariate data can be viewed as coordinates of some multi-dimensional space, in which case one may be interested in some geometric or topological feature of the point cloud data in that space. Methods for determining the "shape" of data are most useful in high-dimensional data, where visual methods cannot be employed. Also, given that identifying qualitative features of data is often the first step of a model-building endeavor, such methods are useful in narrowing down the types of models that are appropriate. The focus of this talk is on one particular method from algebraic topology, called Morse theory, which can reveal some important aspects of the underlying structure of data.

Main Questions

What is the shape of data?

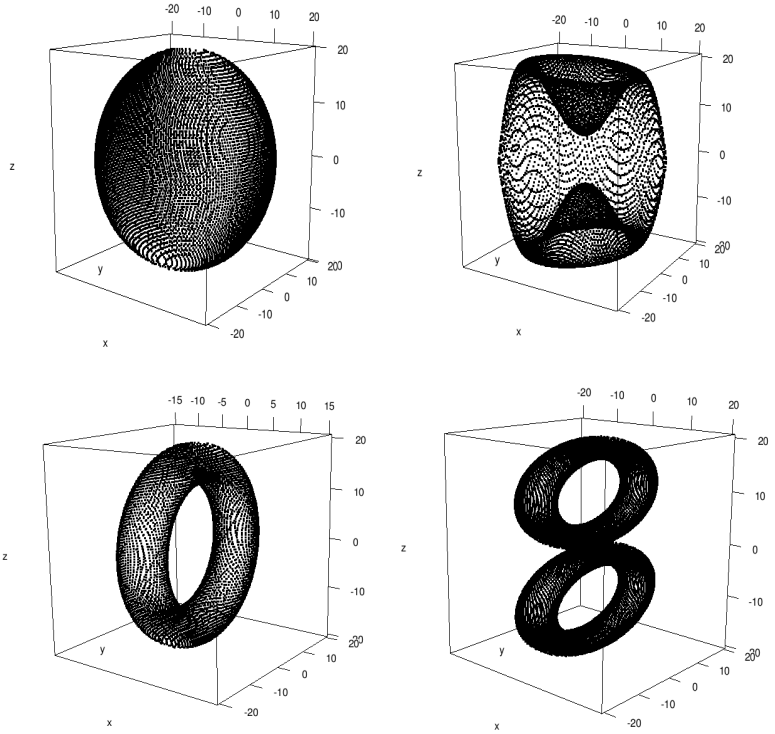
Data: Raw, or transformed by some machine learning model.
Shape: In physical space, or in an abstract space.
Shape: Geometry or **Topology**.

Intro to geometry and topology:

Consider only compact 2d surfaces.

Sphere, perfect, squashed, with or w/o dimple, etc.
are different geometrically,
but same topologically.

genus=0 genus=0



genus=1 genus=2

A sphere, torus, 2-torus, ... are different topologically.

Topological Data Analysis (TDA)

A "new" field.
Same old (useless) questions about identity:
Are NNs same as statistics?
Is TDA same as machine learning?

TDA main contribution:
Data = sample of point cloud data from a manifold in a high-D space.

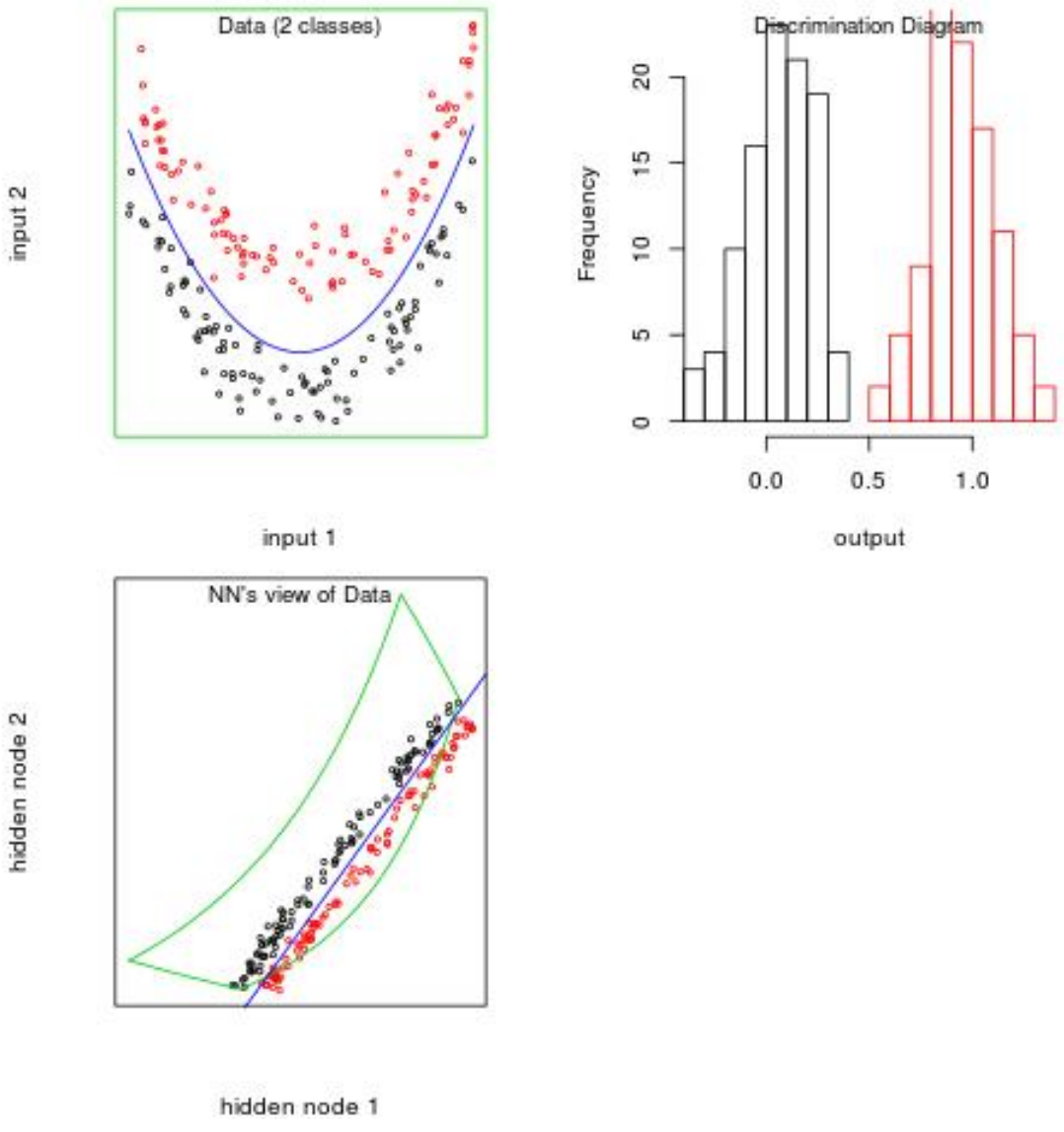
Warning:
Topology has many meanings.
E.g., in character recognition: 0(zero) $\neq$ O (Oh)
There, one talks about topology of graphs.

Why Worry About Topology?

1) Understanding underlying relations in data.
2) Dimensionality reduction. 3) ...

Example 1:

NNs "warp" the feature space.
Especially true in Deep NNs.
Can the warp be interpreted?
Look at the hidden layer as an "image"



See refs for more complicated topology in NN hidden layer.

Example 2: (Thanks to William Hsieh)

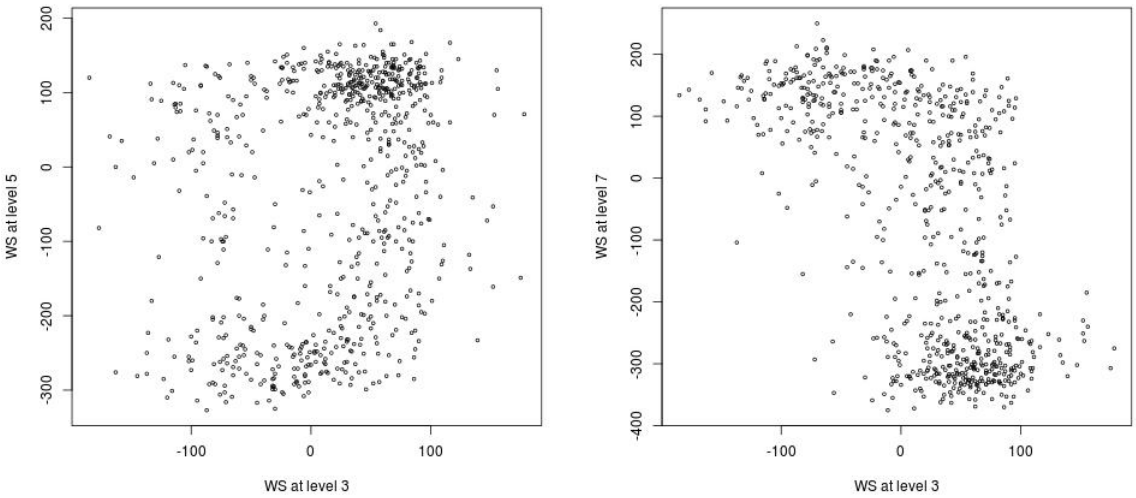
Zonal winds (see refs):

Wind speed at 7 levels: 70 50 40 30 20 15 10 hPaN

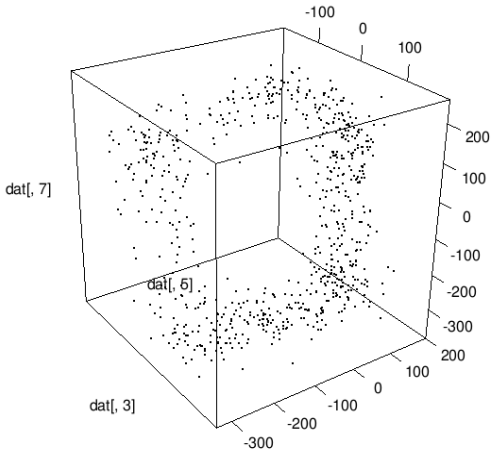
7 is not too big;

can look at 7-choose-2 = 21 scatterplots

But hard to see **big picture** (topology) in 2d:



In 3d, one can see it:



But TDA can reveal the topology without looking.

Useful for understanding and dim reduction.

Topology (Homology)

1) Given a surface S,

Betti numbers are:

b_0 = number of simply-connected components.

b_1 = number of non-contractable loops.

b_2 = number of non-contractable surfaces.

Etc.

Organized into **Poincare polynomial**:

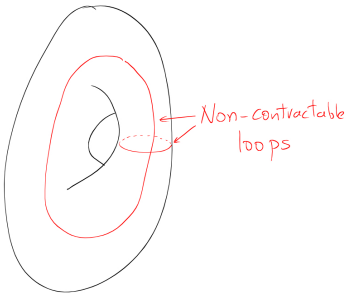
$$P(S) = \sum_k b_k t^k \qquad -1 \le t \le +1.$$

Examples:

$P(\text{Sphere}) = 1 + t^2$

$P(\text{Torus}) = 1 + 2t + t^2$

$P(\text{2-Torus}) = 1 + 4t + t^2$



2) Given a function f on S

$i = 1, 2, \cdots$ critical points, and

n_i = number of non-decreasing directions (Index).

The **Morse polynomial** of the function is

$$M(f) = \sum_i t^{n_i}$$

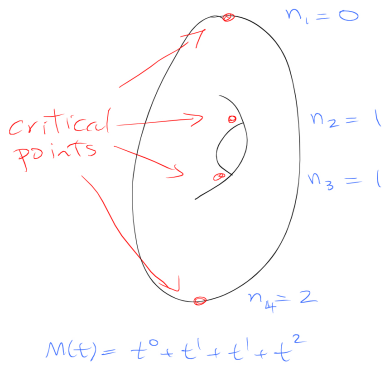
Example: f = height function:

$M(f \text{ on Sphere}) = 1 + t^2$

$M(f \text{ on Dimpled Sphere}) = 2 + 2t + 2t^2$

$M(f \text{ on Torus}) = 1 + 2t + t^2$

$M(f \text{ on 2-Torus}) = 1 + 4t + t^2$



Morse Theory

Morse Inequalities: (Weak and Strong)

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

$Q(t)$ = any polynomial with non-negative coeffs.

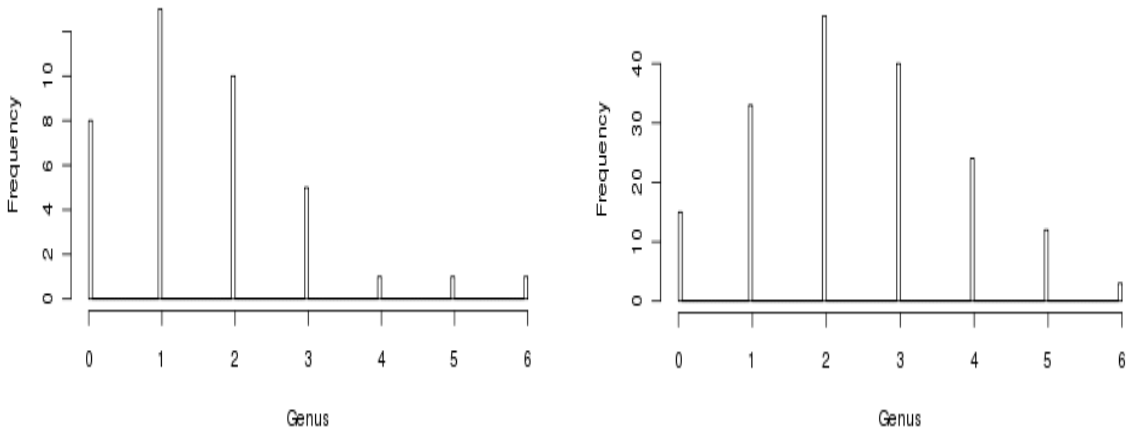
Lacunary principle:
If $M(f) = P(S)$, then f = perfect (defn).
If product of consecutive coeffs in $M(f)$ is 0, then f = perfect.

All together:

$$\mathbf{genus} = \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{2} - \mathbf{n_{min}} + \mathbf{n_{saddle}} - \mathbf{n_{max}})$$

Examples:
Point cloud data on torus and 2-torus.
f = area function (area below any level-surface).
$\rightarrow$ Sampling distribution of genus:

Torus 2-torus



Correctly peaked at 1 and 2.
And we can estimate spread/uncertainty, too.
See refs for more complicated examples.

Conclusion

By simply measuring some function (height, area, ...) one can infer (or put bounds on) the genus (# of handles) of point cloud data.

The genus is important not only for "understanding" the data,
but also for dimensionality reduction.

Afterthought: Please let me know if you have an example that may benefit from these ideas.

References

Marzban, C., and U. Yurtsever 2011: Baby Morse theory in data analysis. Paper at the 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.

Marzban, C., U. Yurtsever, 2011: Baby Morse theory for statistical inference from point cloud data. http://faculty.washington.edu/marzban/morse_data.pdf

http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Zonal winds: http://www.geo.fu-berlin.de/met/ag/strat/produkte/qbo/qbo.dat
(Thanks to William Hsieh for suggesting this data set)