TOPOLOGICAL DATA ANALYSIS

Fernando Schwartz
Mathematics Department
UTK

“In God we trust; all others must bring data.”
W.E. Deming
WHAT IS Persistent Homology?

Shmuel Weinberger

In memory of my friend, Partha Niyogi (1967–2010)

Consider the art of Seurat or a piece of old newsprint. The eye, or the brain, performs the marvelous task of taking the sense data of individual points and assembling them into a coherent image of a continuum—it infers the continuous from the discrete.

Difficult issues of a similar sort occur in many problems of data analysis. One might have samples that are chosen nonuniformly (e.g., not filling a grid), and, moreover, one is constantly plagued by problems of noise—the data can be corrupted in various ways.

Pure mathematicians have problems of this sort as well. One is often interested in inferring properties of an enveloping space from a discrete object within it or, in reverse, seeking commonalities of all the discrete subobjects of a given continuous one. To give one example, this theme is a central one in geometric group theory, in which a typical problem, going back to Furstenberg and Mostow, asks to reconstruct a connected Lie group from a lattice in it.

And, ubiquitously in analysis, one often tries to get information about a function from approximations to it. For instance, any function uniformly close to \( z \rightarrow z^n \) on the complex plane necessarily has at least \( n \) roots (with multiplicity).

Because topology is essentially a qualitative field, it is perhaps not surprising that there has been a development of some common topological technology for these problems. Needless to say, Shmuel Weinberger is professor of mathematics at the University of Chicago. His email address is shmuel@math.uchicago.edu.

The author was partially supported by NSF 0805913, NSF-SGER grant DMS-0852227, and STOMP, a DARPA program.

The Seine at La Grande Jatte by Georges Seurat.

The deepest aspects of these problems all have the idiosyncrasies of their particular application domains. The focus here is on what is common to them.

For simplicity in what follows, we will take all homology groups to have coefficients in a field.

Definition. Suppose that we have \( \{X_r\} \) an estimated sequence of spaces (satisfying mild technical conditions), parameterized by the real numbers. We define the \( k \)th persistent homology \( \text{PH}_k(X) \) by the formula:

\[
\text{PH}_k(X) = \Pi \text{H}_k(X_r).
\]

The product on the right is an awful object: formally it is an uncountable dimensional vector space, but there is a reasonable way to make sense of this, taking into account the fact that...
Coffee Mug is topologically equivalent to a Donut
Image from DARPA. Calls for advances in “big data” to help the warfighter
3 APPLICATIONS OF TDA

FROM: LUM, P.Y. ET AL. EXTRACTING INSIGHTS FROM THE SHAPE OF COMPLEX DATA USING TOPOLOGY. SCI. REP. 3, 1236; DOI:10.1038/SREP01236 (2013)

- Breast cancer sub-population with excellent survival
- Networks of the US House of Representatives based on voting behavior
- Basketball team stratification
Breast cancer patients with the same stage of disease can have different treatment responses and overall outcome.

The strongest predictors for metastases fail to classify accurately breast tumors according to their clinical behavior.

Chemotherapy or hormonal therapy reduces the risk of distant metastases by about one-third; however, 70-80% of patients receiving this treatment would have survived without it.

Supervised Classification of DNA microarray analysis on primary breast tumors was used to identify a gene signature strongly predictive of a short interval to distant metastases (poor prognosis) in patients without tumor cells in local lymph nodes at diagnosis.

Signatures of breast cancer gene expression allow for patient-tailored therapy strategies.
CANCER DATA: 5,000 GENES (CLUSTERING METHOD)
Breast cancer prognosis can be derived from the gene expression profile of the primary tumor.

Up to 90% of lymph-node-negative young breast cancer patients are candidates for adjuvant systemic treatment. Nevertheless, 70-80% of these patients would not have developed distant metastases without adjuvant treatment; these patients may not benefit from the treatment, and may potentially suffer from the side effects.

The prognostic profile provides a powerful tool to tailor adjuvant systemic treatment that could greatly reduce the cost of breast cancer treatment, both in terms of adverse side effects and health care expenditure. Genes that are over-expressed in tumors with a poor prognosis profile are potential targets for the rational development of new cancer drugs.

Identification of such targets may improve the efficiency of developing therapeutics for many tumor types.
Breast cancer continues to confound us, with multiple sub-types identified over the years. Many researchers have continued to find patient sub-populations that may be important for targeted therapy. The coordinate invariance and the related insensitivity to deformation are useful in reconciling the results from two distinct data sets. We show here that topological maps can more finely stratify patients with breast cancer.

Identifying patient subsets in breast cancer

We use two relatively older breast cancer data sets, NKI and GSE2034, to demonstrate that even with older data sets, there is much to be gained by using this approach. The first data set, NKI, consists of gene expression levels extracted from 272 tumors and is analyzed using about 1500 most varying genes. Although we are able to compute with any variance threshold, we show in Figure S1 that the shape in the network becomes less distinct as the threshold is relaxed. We therefore used the top most varying genes for this example. In addition to gene expression data, the NKI data set includes survival information. Figure 2 (panel A) shows the network constructed using both gene expression columns and survival norm. The survival filter is used as a supervised step to study the behavior of the survivors separately from the non-survivors (filter functions described in Materials and Methods).

An important parameter to the TDA implementation is the choice of filters. Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn’t survive, the bottom half are the patient sub-groups that may be important for targeted therapy.

In order to show the applications of TDA in the real world, we used the NKI data set to explore patient survivors and non-survivors with low ESR1 expression levels. ESR1 levels are strongly correlated with poor prognosis although not all patients with low ESR1 levels have poor survival. It is also generally understood that low ESR1 levels often correspond to poor prognosis. However, it is also possible for patients with low ESR1 levels but high survival to be identified over the years. Many researchers have continued to find molecular sub-types of breast cancer. Note that the resulting network has a structure shaped like a straight line. One works with a data set experimentally to find values that go into the analyses, as well as the definition of filters related to the singular value decomposition, in the Methods section. The innovation in this paper is to demonstrate that even with older data sets, there is much to be gained by using this approach.

To demonstrate this, we applied TDA to two data sets. The first data set, NKI, consists of gene expression levels extracted from 272 tumors and is analyzed using about 1500 most varying genes. Although we are able to compute with any variance threshold, we show in Figure S1 that the shape in the network becomes less distinct as the threshold is relaxed. We therefore used the top most varying genes for this example. In addition to gene expression data, the NKI data set includes survival information. Figure 2 (panel A) shows the network constructed using both gene expression columns and survival norm. The survival filter is used as a supervised step to study the behavior of the survivors separately from the non-survivors (filter functions described in Materials and Methods).

An important parameter to the TDA implementation is the choice of filters. Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn’t survive, the bottom half are the patient sub-populations in breast cancer. We show here that topological maps can more finely stratify patients with breast cancer.

Identifying patient subsets in breast cancer

We use two relatively older breast cancer data sets, NKI and GSE2034, to demonstrate that even with older data sets, there is much to be gained by using this approach. The first data set, NKI, consists of gene expression levels extracted from 272 tumors and is analyzed using about 1500 most varying genes. Although we are able to compute with any variance threshold, we show in Figure S1 that the shape in the network becomes less distinct as the threshold is relaxed. We therefore used the top most varying genes for this example. In addition to gene expression data, the NKI data set includes survival information. Figure 2 (panel A) shows the network constructed using both gene expression columns and survival norm. The survival filter is used as a supervised step to study the behavior of the survivors separately from the non-survivors (filter functions described in Materials and Methods).

An important parameter to the TDA implementation is the choice of filters. Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn’t survive, the bottom half are the patient sub-populations in breast cancer. We show here that topological maps can more finely stratify patients with breast cancer.

Identifying patient subsets in breast cancer

We use two relatively older breast cancer data sets, NKI and GSE2034, to demonstrate that even with older data sets, there is much to be gained by using this approach. The first data set, NKI, consists of gene expression levels extracted from 272 tumors and is analyzed using about 1500 most varying genes. Although we are able to compute with any variance threshold, we show in Figure S1 that the shape in the network becomes less distinct as the threshold is relaxed. We therefore used the top most varying genes for this example. In addition to gene expression data, the NKI data set includes survival information. Figure 2 (panel A) shows the network constructed using both gene expression columns and survival norm. The survival filter is used as a supervised step to study the behavior of the survivors separately from the non-survivors (filter functions described in Materials and Methods).

An important parameter to the TDA implementation is the choice of filters. Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn’t survive, the bottom half are the patient sub-populations in breast cancer. We show here that topological maps can more finely stratify patients with breast cancer.

Identifying patient subsets in breast cancer

We use two relatively older breast cancer data sets, NKI and GSE2034, to demonstrate that even with older data sets, there is much to be gained by using this approach. The first data set, NKI, consists of gene expression levels extracted from 272 tumors and is analyzed using about 1500 most varying genes. Although we are able to compute with any variance threshold, we show in Figure S1 that the shape in the network becomes less distinct as the threshold is relaxed. We therefore used the top most varying genes for this example. In addition to gene expression data, the NKI data set includes survival information. Figure 2 (panel A) shows the network constructed using both gene expression columns and survival norm. The survival filter is used as a supervised step to study the behavior of the survivors separately from the non-survivors (filter functions described in Materials and Methods).

An important parameter to the TDA implementation is the choice of filters. Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn’t survive, the bottom half are the patient sub-populations in breast cancer. We show here that topological maps can more finely stratify patients with breast cancer.
Recognizing shapes or patterns is critical to discovering insights in the data and identifying meaningful sub-groups.

Typical features are FLARES and LOOPS.

Topological methods act as geometric approach to pattern or shape recognition within data.
HOW DOES TDA COMPARE?

In red: new TDA sub-group
VOTING BEHAVIOR OF THE US HOUSE OF REPRESENTATIVES

<table>
<thead>
<tr>
<th>Year</th>
<th>Principal SVD value</th>
<th>Secondary SVD value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td><img src="image1" alt="2009 Scatter Plot" /></td>
<td><img src="image2" alt="2009 Scatter Plot" /></td>
</tr>
<tr>
<td>2010</td>
<td><img src="image3" alt="2010 Scatter Plot" /></td>
<td><img src="image4" alt="2010 Scatter Plot" /></td>
</tr>
</tbody>
</table>

PCA identifies 2 groups
This classification has been refined further into five distinct positions, namely point guard, shooting guard, small forward, power forward, and center. These positions represent a spectrum of players from short, fast, and playing outside the key to tall, slow, and playing inside the key. However, distinguishing players based only on physical characteristics such as height or speed is perhaps arbitrary and outdated. One can then ask the question if there is instead a more informative stratification of player types based on their in-game performance. To answer this question, we constructed performance profiles for each of the 452 players in the NBA by using data from the 2010–2011 NBA season (Figure 5).

From the networks, we see a much finer structure than five distinct categories. These structures represent groups of players based on their in-game performance statistics. For example, the left side of the main network reveals a finer stratification of guards into three groups, namely offensive point guards, defensive point guards, and ball handling shooting guards. We also see three smaller structures in the lower central part of the map that we labeled “All NBA” and “All NBA 2nd team”. The “All NBA” network consists of the NBA’s most exceptional players and the second team consists of players who are also all-around excellent players but perhaps not as top-performing as the “All NBA” players. Within “All NBA” group are all-star players like LeBron James and Kobe Bryant. Interestingly, there are some less well-known players in the “All NBA” network such as Brook Lopez, suggesting that they are potential up and coming stars.

It is of note that the “All NBA” and “All NBA 2nd team” networks are well separated from the large network, indicating that their in-game statistics are very different. To also illustrate the capability to perform multi-resolution analyses simultaneously on the same dataset and how that kind of analysis is important, we compared the high resolution network (Figure 5, right panel) to the lower resolution network (Figure 5, left panel). The right panel shows that at a lower resolution, these players form 4 categories, which are scoring big men, paint protectors, scoring guards, and ball handling guards. In summary, this topological network suggests a much finer stratification of players into thirteen positions rather than the traditional division into five positions.

Discussion

We have shown that TDA can handle a variety of data types using three real world examples. The three key concepts of topological methods, coordinate freeness, invariance to deformation and compressed representations of shapes are of particular value for applications to data analysis. Coordinate free analysis means that the representation is independent of the particular way in which the data set is given coordinates, but rather depends only on the similarity of the points as reflected in the distance function. Coordinate free representations are vital when one is studying data collected with different technologies, or from different labs when the methodologies cannot be standardized. The invariance to deformation provides some robustness to noise. Compressed representations are obviously important when one is dealing with very large data sets, but even...
FRAGMENTATION INDEX LAST 20 YRS.

This classification has been refined further into five distinct positions, namely point guard, shooting guard, small forward, power forward, and center. These positions represent a spectrum of players from short, fast, and playing outside the key to tall, slow, and playing inside the key. However, distinguishing players based only on physical characteristics such as height or speed is perhaps arbitrary and outdated. One can then ask the question if there is instead a more informative stratification of player types based on their in-game performance.

To answer this question, we constructed performance profiles for each of the 452 players in the NBA by using data from the 2010–2011 NBA season.

From the networks, we see a much finer structure than five distinct categories. These structures represent groups of players based on their in-game performance statistics. For example, the left side of the main network reveals a finer stratification of guards into three groups, namely offensive point guards, defensive point guards, and ball handling shooting guards. We also see three smaller structures in the lower central part of the map that we labeled “All NBA” and “All NBA 2nd team”. The “All NBA” network consists of the NBA’s most exceptional players and the second team consists of players who are also all-around excellent players but perhaps not as top-performing as the “All NBA” players. Within “All NBA” group are all-star players like LeBron James and Kobe Bryant. Interestingly, there are some less well-known players in the “All NBA” network such as Brook Lopez, suggesting that they are potential up and coming stars.

It is of note that the “All NBA” and “All NBA 2nd team” networks are well separated from the large network, indicating that their in-game statistics are very different. To also illustrate the capability to perform multi-resolution analyses simultaneously on the same dataset and how that kind of analysis is important, we compared the high resolution network to the lower resolution network.

In summary, this topological network suggests a much finer stratification of players into thirteen positions rather than the traditional division into five positions.

Discussion

We have shown that TDA can handle a variety of data types using three real world examples. The three key concepts of topological methods, coordinate freeness, invariance to deformation and compressed representations of shapes are of particular value for applications to data analysis. Coordinate free analysis means that the representation is independent of the particular way in which the data set is given coordinates, but rather depends only on the similarity of the points as reflected in the distance function. Coordinate free representations are vital when one is studying data collected with different technologies, or from different labs when the methodologies cannot be standardized. The invariance to deformation provides some robustness to noise. Compressed representations are obviously important when one is dealing with very large data sets, but even

---

**Figure 4**

Top panel is the fragmentation index calculated from the number of sub-networks formed each year per political party. X-axis: 1990–2011. Y-axis: Fragmentation index. Color bars denote, from top to bottom, party of the President, party for the House, party for the Senate (red: republican; blue: democrat; purple: split). The bottom 3 panels are the actual topological networks for the members. Networks are constructed from voting behavior of the member of the house, with an “aye” vote coded as a 1, “abstain” as zero, and “nay” as a -1. Each node contains sets of members. Each panel labeled with the year contains networks constructed from all the members for all the votes of that year. Note high fragmentation in 2010 in both middle panel and in the Fragmentation Index plot (black bar). The distance metric and filters used in the analysis were Pearson correlation and principal and secondary metric SVD. Metric: Correlation; Lens: Principal SVD Value (Resolution 120, Gain 4.5x, Equalized) and Secondary SVD Value (Resolution 120, Gain 4.5x, Equalized). Color: Red: Republican; Blue: Democrats.

**Fragmentation index:** number of connected components
FRAGMENTATION PEAKS

2008: Economic meltdown

2010: Obamacare
2009 GOP SPLIT

Top issues that caused divisions

“CLASSIC” BASKETBALL TEAM STRATIFICATION

- Point guard
- Shooting guard
- Small forward
- Power forward
- Center

- Classification based on physical characteristics, such as height or speed.
- Outdated?
Data: performance profiles of all 452 NBA players from 2010-2011 season.

Low-Resolution network reveals 4 categories:

- Scoring Big
- Paint protectors
- Scoring guards
- Ball handling guards
HIGHER RES: 13 POSITIONS!

All star NBA: LeBron James, Kobe Bryant --but also Brook Lopez (up and coming star?)
NEW STRATIFICATION

- Offensive Ball-Handler: This guy handles the ball and specializes in points, free throws and shots attempted, but is below average in steals and blocks.
- Defensive Ball-Handler: This is a defense-minded player who handles the ball and specializes in assists and steals, but is only so-so when it comes to points, free throws and shots.
- Combo Ball-Handler: These players are adept at both offense and defense but don’t stand out in either category.
- Shooting Ball-Handler: Someone with a knack for scoring, characterized by above-average field goal attempts and points.
- Role-Playing Ball-Handler: These guys play fewer minutes and don’t have as big a statistical impact on the game.
- 3-Point Rebounder: Such a player is a ball-handler and big man above average in rebounds and three-pointers, both attempted and made, compared to ball-handlers.
- Scoring Rebounder: He grabs the ball frequently and demands attention when on offense.
- Paint Protector: A big man like Marcus Camby and Tyson Chandler known for blocking shots and getting rebounds, but also for racking up more fouls than points.
- Scoring Paint Protector: These players stand out on offense and defense, scoring, rebounding and blocking shots at a very high rate.
- NBA 1st-Team: This is a select group of players so far above average in every statistical category that the software simply groups them together regardless of their height or weight. Kevin Durant and LeBron James fall in this category.
- NBA 2nd-Team: Not quite as good, but still really, really good. Rudy Gay and Caron Butler are examples.
- Role Player: Slightly less skilled than the 2nd-team guys, and they don’t play many minutes. Guys like Shane Battier and Ronnie Brewer fall under this position.
- One-of-a-Kind: These guys are so good they are off the charts. The software could not connect them to any other player. Derrick Rose and Dwight Howard are examples.
TOPOLOGICAL DATA ANALYSIS
DISCUSSION

TDA can handle a variety of data types.

3 key concepts of Topological Methods:

- Coordinate Free (metric spaces)
- Deformation Invariance (coffee mug “=” donut)
- Compressed Representation (simplicial complex)
COORDINATE FREE REPRESENTATIONS

- Topology builds up using a METRIC SPACE, which is a SET with a notion of DISTANCE between any pair of POINTS.

- Distance functions specifies the SHAPE of an object independent of its “position.”

- By using an approach that only depends on distances (coordinate free), topology can compare data derived from different platforms (coordinate systems).
DEFORMATION INVARIANCE

- By STRETCHING A SHAPE we make small deformations which do not affect the topology. For example, stretching does not change the amount of LOOPS and HOLES present. (In this world, coffee cup has the same “shape” as a donut)

- DEFORMATION INVARIANT notions capture the object’s most fundamental SHAPE PROPERTIES. The approach possesses the ability to PICK OUT THE “SHAPE” of an object despite variations.

- This helps the analysis be LESS SENSITIVE TO NOISE.
COMPRESSED REPRESENTATIONS

- INFINITELY COMPLEX SHAPES are TOPOLOGICALLY EQUIVALENT to polygonal representations, or SIMPLICIAL COMPLEXES, which are encoded using FINITE amounts of data.

- This is achieved via computing a TRIANGULATION of the object.

- This can be regarded as a form of COMPRESSION, where the number of points goes from INFINITE to FINITE.

- Some information, such as curvature, is LOST in the process.
BOTTOM LINE

- TDA sensitive to both LARGE and SMALL SCALE patterns that fail to be detected by other methods such as Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and Cluster Analysis.

- PCA and MDS produce unstructured scatterplots and Clustering Methods produce distinct, unrelated groups.

- Standard methodologies sometimes obscure geometric features that topological methods capture.
WANT TO KNOW MORE?


MAIN (MATH) INGREDIENTS

- Graphs
- Surfaces (2-dimensional manifolds)
- Triangulations
- Simplicial Complexes
- Homology Groups
- Duality (cohomology)
- Morse functions
- Persistent Homology
- Stability Theorems
“A sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing the radius, holes appear and disappear. Which holes are real and which are noise?”