Statistics for a Computational Topologist

Part I

Brittany Terese Fasy
TA: Samuel A. Micka

School of Computing and Dept. of Mathematical Sciences
Montana State University

August 14, 2018
Why Topological Data Analysis?

“Data has shape and the shape matters.” - Gunnar Carlsson

Today, Data is high-dimensional,

Nicolaua, Levine, and Carlsson, PNAS 2011

... and needs to be summarized, analyzed, and compared!
Why Topological Data Analysis?

“Data has shape and the shape matters.” - Gunnar Carlsson

Today, Data is high-dimensional, HUGE,

Nicolaua, Levine, and Carlsson, PNAS 2011

http://astrobites.com/

... and needs to be summarized, analyzed, and compared!
Why Topological Data Analysis?

“Data has shape and the shape matters.” - Gunnar Carlsson

Today, Data is high-dimensional, HUGE, present everywhere

... and needs to be summarized, analyzed, and compared!

Nicolaua, Levine, and Carlsson, PNAS 2011

http://astrobites.com/

www.mapconstruction.org
What questions do we ask in data analysis?
What questions do we ask in data analysis?

- Think! Write down one question (2 min)

Pair! Share with partner, and add more questions to your list (5 min)

Share! Raise hands please!

More ideas?

mickas37@gmail.com
What questions do we ask in data analysis?

- Think! Write down one question (2 min)
- Pair! Share with partner, and add more questions to your list (5 min)
What questions do we ask in data analysis?

- Think! Write down one question (2 min)
- Pair! Share with partner, and add more questions to your list (5 min)
- Share! Raise hands please! (5 min)
What questions do we ask in data analysis?

- Think! Write down one question (2 min)
- Pair! Share with partner, and add more questions to your list (5 min)
- Share! Raise hands please! (5 min)
- More ideas? mickas37@gmail.com
Summarize and Analyze

- What is this shape?
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare

- Are these the same?
Data Analysis Questions

Summarize and Analyze
- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare
- Are these the same? In distribution?
Data Analysis Questions

Summarize and Analyze
- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare
- Are these the same? In distribution?
- Has something changed?
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare

- Are these the same? In distribution?
- Has something changed? If so, what has changed?
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare

- Are these the same? In distribution?
- Has something changed? If so, what has changed?
- Which is bigger?
Data Analysis Questions

Summarize and Analyze
- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare
- Are these the same? In distribution?
- Has something changed? If so, what has changed?
- Which is bigger?
- Can we retain the null hypothesis? (Inference: Hypothesis Testing)
Data Analysis Questions

Summarize and Analyze

- What is this shape? How many components / populations?
- Can we categorize? (Classification)
- What are the parameters? (Inference: Point Estimation)
- How far do parameters likely lie from estimates? (Confidence Sets)

Compare

- Are these the same? In distribution?
- Has something changed? If so, what has changed?
- Which is bigger?
- Can we retain the null hypothesis? (Inference: Hypothesis Testing)
- What is the relationship between $X$ and $Y$? (Regression)
Most Important Questions

1. Which descriptor best captures our data?

2. How do we measure distance between descriptors?
1. Which descriptor best captures our data?
1. Which descriptor best captures our data?
   - Descriptors
   - Confidence Sets
Most Important Questions

1. Which descriptor best captures our data?
   - Descriptors
   - Confidence Sets

2. How do we measure distance between descriptors?
Most Important Questions

1. Which descriptor best captures our data?
   - Descriptors
   - Confidence Sets

2. How do we measure distance between descriptors?
   - Distances
   - Clustering
Topological Descriptors
Stat Reverences


Let $F$ be a probability distribution with density $f$. 

$X \sim F$ reads “$X$ has distribution $F$”. Here, $X$ is called a random variable.

Expectation: $E(X) = \int x \, dF(x)$. 

Quantile Function: CDF$^{-1}(q)$. 

![Graphs of two distributions](image-url)
Stat Slide: The Basics

- Let $F$ be a probability distribution with density $f$.
- $X \sim F$ reads “$X$ has distribution $F$”.
- Here, $X$ is called a random variable.
Let $F$ be a probability distribution with density $f$.

$X \sim F$ reads “$X$ has distribution $F$”.

Here, $X$ is called a *random variable*.

Expectation: $\mathbb{E}(X) = \int x \, dF(x)$.
Let $F$ be a probability distribution with density $f$.

$X \sim F$ reads “$X$ has distribution $F$”.

Here, $X$ is called a random variable.

Expectation: $\mathbb{E}(X) = \int x \, dF(x)$.

Quantile Function $CDF^{-1}(q)$. 

![Diagram illustrating probability distribution and quantile function]
Let $F$ be some distribution.

Let $X_1, X_2, \ldots, X_n \sim F$. (The data).
Let $F$ be some distribution.

Let $X_1, X_2, \ldots, X_n \sim F$. (The data).

A statistic or descriptor is a function of the data:

$T(X_1, X_2, \ldots, X_n)$ or $T(X^n)$.
Let $F$ be some distribution.

Let $X_1, X_2, \ldots, X_n \sim F$. (The data).

A statistic or descriptor is a function of the data: $T(X_1, X_2, \ldots, X_n)$ or $T(X^n)$.

Sample average: $\bar{X}^n = \frac{1}{n} \sum X_i$. 

Law of Large Numbers

$\bar{X}^n$ converges to $E(X_i)$ in probability: $\forall \epsilon > 0, \lim_{n \to \infty} (|P(\bar{X}^n - E(X_i))| > \epsilon) \to 0$.

Central Limit Theorem

$\sqrt{n}(\bar{X}^n - E(X_i))$ converges in distribution to a Normal distribution, i.e., sample average is approximately Normal for large enough samples.
Let $F$ be some distribution.

Let $X_1, X_2, \ldots, X_n \sim F$. (The data).

A statistic or descriptor is a function of the data: $T(X_1, X_2, \ldots, X_n)$ or $T(X^n)$.

Sample average: $\bar{X}^n = \frac{1}{n} \sum X_i$.

Law of Large Numbers

$\bar{X}^n$ converges to $\mathbb{E}(X_i)$ in probability:
Let $F$ be some distribution.

Let $X_1, X_2, \ldots, X_n \sim F$. (The data).

A statistic or descriptor is a function of the data:

$$T(X_1, X_2, \ldots, X_n) \text{ or } T(X^n).$$

Sample average: $\bar{X}^n = \frac{1}{n} \sum X_i$.

**Law of Large Numbers**

$\bar{X}^n$ converges to $\mathbb{E}(X_i)$ in probability:

$$\forall \varepsilon > 0, \lim_{n \to \infty} \left( \mathbb{P}(\bar{X}^n - \mathbb{E}(X_i) > \varepsilon) \right) \to 0.$$
Prob/Stat Slide: Descriptors and Limit Theory

- Let $F$ be some distribution.
- Let $X_1, X_2, \ldots, X_n \sim F$. (The data).
- A statistic or descriptor is a function of the data: $T(X_1, X_2, \ldots, X_n)$ or $T(X^n)$.
- Sample average: $\bar{X}^n = \frac{1}{n} \sum X_i$.

**Law of Large Numbers**

$\bar{X}^n$ converges to $\mathbb{E}(X_i)$ in probability:

$$\forall \varepsilon > 0, \lim_{n \to \infty} (\mathbb{P}(\bar{X}^n - \mathbb{E}(X_i) > \varepsilon)) \to 0.$$

**Central Limit Theorem**

$\sqrt{n}(\bar{X}^n - \mathbb{E}(X_i))$ converges in distribution to a Normal distribution,
Let $F$ be some distribution.

Let $X_1, X_2, \ldots, X_n \sim F$. (The data).

A statistic or descriptor is a function of the data:

$T(X_1, X_2, \ldots, X_n)$ or $T(X^n)$.

Sample average: $\bar{X}^n = \frac{1}{n} \sum X_i$.

Law of Large Numbers

$\bar{X}^n$ converges to $\mathbb{E}(X_i)$ in probability:

$$\forall \varepsilon > 0, \lim_{n \to \infty} (P(\bar{X}^n - \mathbb{E}(X_i) > \varepsilon)) \to 0.$$ 

Central Limit Theorem

$\sqrt{n}(\bar{X}^n - \mathbb{E}(X_i))$ converges in distribution to a Normal distribution, i.e., sample average is approximately Normal for large enough samples.
Data as Point Clouds
Descriptive Image Captions:
- Noise
- Big Loop
- Pinch

Diagram Description:
The diagram illustrates the progression from noise to big loop to pinch. It shows a series of images representing different stages in a data analysis process, emphasizing the transition from noisy data to more structured data forms, culminating in the identification of topological features like loops and pinches.
Data as Persistence Diagrams
Confidence Sets for Persistence Diagrams:
Analyzing Descriptors
Objective

To Find a Threshold

Given $\alpha \in (0, 1)$, we will find $q^\alpha > 0$ such that

$$\mathbb{P}(W_\infty(D, \hat{D}_n) \leq q^\alpha) \geq 1 - \alpha.$$
Objective

To Find a Threshold

Given $\alpha \in (0, 1)$, we will find $q^\alpha > 0$ such that

$$P(W_\infty(D, \hat{D}_n) \leq q^\alpha) \geq 1 - \alpha.$$

References

- Chazal, BTF, Lecci, Michel, Rinaldo, and Wasserman. Robust Topological Inference: Distance To a Measure and Kernel Distance, JMLR 18(159):1–40, 2018.
Old idiom: “pull yourself up by your bootstraps”
Stat Slide: Bootstrapping

Old idiom: “pull yourself up by your bootstraps”
Want: a parameter of an unknown distribution $F$. 
Old idiom: “pull yourself up by your bootstraps”
Want: a parameter of an unknown distribution $F$.
Try: estimate using empirical distribution $\hat{F}$. 
Old idiom: “pull yourself up by your bootstraps”
Want: a parameter of an unknown distribution $F$.
Try: estimate using empirical distribution $\hat{F}$.
Nonparametric technique!
Bottleneck Bootstrap

We have a point cloud sample:
\[ S_n = \{ X_1, \ldots, X_n \}; \ X_i \sim P. \]
Bottleneck Bootstrap

We have a point cloud sample: \( S_n = \{X_1, \ldots, X_n\} \); \( X_i \sim P \).

Subsample (with replacement), obtaining: \( X = \{X_1^*, \ldots, X_b^*\} \)
**Bottleneck Bootstrap**

We have a point cloud sample: 
\[ S_n = \{X_1, \ldots, X_n\}; \quad X_i \sim P. \]

Subsample (with replacement), obtaining: 
\[ X = \{X_1^*, \ldots, X_b^*\} \]

Compute 
\[ \hat{\Theta}_b^*(X^*) = W_\infty(X^*, S_n) \]
using KDE or DTM.
Bottleneck Bootstrap

We have a point cloud sample: $S_n = \{X_1, \ldots, X_n\}$; $X_i \sim P$.

Subsample (with replacement), obtaining: $X = \{X^*_1, \ldots, X^*_b\}$

Compute $\hat{\Theta}^*_b(X^*) = W_{\infty}(X^*, S_n)$ using KDE or DTM.

Consider all possible outcomes:

$\{\hat{\Theta}^*_b(X^*)\}_{X^* \subset S_n}$
Confidence Sets

Bottleneck Bootstrap

We have a point cloud sample: 
\[ S_n = \{X_1, \ldots, X_n\}; \ X_i \sim P. \]

Subsample (with replacement), obtaining: 
\[ X = \{X^*_1, \ldots, X^*_b\} \]

Compute \( \hat{\Theta}_b(X^*) = W_{\infty}(X^*, S_n) \) using KDE or DTM.

Consider all possible outcomes:
\[ \{\hat{\Theta}_b(X^*)\} \ x^* \subset S_n \]

Mimics:
\[ \{\Theta(X) = W_{\infty}(S_n, M)\} \ x^* \subset S_n \subset M \]
We have a point cloud sample: $S_n = \{X_1, \ldots, X_n\}; \ X_i \sim P$.

Subsample (with replacement), obtaining: $X = \{X_1^*, \ldots, X_b^*\}$

Compute $\hat{\Theta}_b^*(X^*) = W_\infty(X^*, S_n)$ using KDE or DTM.

Consider all possible outcomes:

$$\{\hat{\Theta}_b^*(X^*)\}_{X^* \subset S_n}$$

**Mimics:**

$$\{\Theta(X) = W_\infty(S_n, M)\}_{S_n \subset M}$$
Confidence Sets for Persistent Diagrams

\[ C_\alpha = \{ D \in \mathcal{D}_T : W_\infty(D, \hat{D}_n) \leq q^\alpha \} \]
Confidence Sets for Persistent Diagrams

\[ C_\alpha = \{ D \in \mathcal{D}_T : W_\infty(D, \hat{D}_n) \leq q_\alpha \} \]
Example
Challenges
Challenges

Techniques

- Prove limit theorems.
Challenges

Techniques

- Prove limit theorems.
- Determine suitable assumptions on input.
Challenges

Techniques

- Prove limit theorems.
- Determine suitable assumptions on input.
- Use the geometry of input (e.g., properties of an underlying smooth manifold).
Challenges

Techniques

- Prove limit theorems.
- Determine suitable assumptions on input.
- Use the geometry of input (e.g., properties of an underlying smooth manifold).

Questions

- These results are *in the limit*. When is $n$ big enough?
Challenges

Techniques
- Prove limit theorems.
- Determine suitable assumptions on input.
- Use the geometry of input (e.g., properties of an underlying smooth manifold).

Questions
- These results are *in the limit*. When is $n$ big enough?
- What confidence sets can we construct in the multi-d setting?
Challenges

Techniques

- Prove limit theorems.
- Determine suitable assumptions on input.
- Use the geometry of input (e.g., properties of an underlying smooth manifold).

Questions

- These results are in the limit. When is $n$ big enough?
- What confidence sets can we construct in the multi-d setting?
- What is the optimal threshold for particular filtrations?
Challenges

Techniques
- Prove limit theorems.
- Determine suitable assumptions on input.
- Use the geometry of input (e.g., properties of an underlying smooth manifold).

Questions
- These results are in the limit. When is $n$ big enough?
- What confidence sets can we construct in the multi-d setting?
- What is the optimal threshold for particular filtrations?
- Power analysis: are the rejected points topologically insignificant? (Type II errors)
Confidence Sets

Challenges

Techniques

- Prove limit theorems.
- Determine suitable assumptions on input.
- Use the geometry of input (e.g., properties of an underlying smooth manifold).

Questions

- These results are in the limit. When is $n$ big enough?
- What confidence sets can we construct in the multi-d setting?
- What is the optimal threshold for particular filtrations?
- Power analysis: are the rejected points topologically insignificant? (Type II errors)
Distance Measures:
Comparing Descriptors
Distances

= ?

B. Fasy (MSU)  Statistics for a Computational Topologist  August 14, 2018

20 / 25
Distances Between Diagrams

- Bottleneck $d_\infty$.
- Interleaving distance.
- Wasserstein $d_p$.
- Erosion distance.


Distances Between Diagrams

- Bottleneck $d_{\infty}$.
- Interleaving distance.
- Wasserstein $d_p$.
- Erosion distance.

**Question**

Can we define a centroid / Fréchet mean?

$$\arg \min_D \sum_i W_2^2(D, D_i)$$
Distances Between Diagrams

- Bottleneck $d_\infty$.
- Interleaving distance.
- Wasserstein $d_p$.
- Erosion distance.

**Question**

Can we define a centroid / Fréchet mean?

$$\arg\min_D \sum_i W_\infty^2(D, D_i)$$

Distances Between Diagrams

- Bottleneck $d_\infty$.
- Interleaving distance.
- Wasserstein $d_p$.
- Erosion distance.

**Question**
Can we define a centroid / Fréchet mean?

$$\arg \min_D \sum_i W_\infty^2(D, D_i)$$

Clustering
Clustering

Clustering (Unsupervised Learning)
Clustering (Unsupervised Learning)

- Hierarchical: agglomerative or divisive.
Clustering

Clustering (Unsupervised Learning)

- Hierarchical: agglomerative or divisive.
- $k$-means: NP-hard, so algorithms find a local minimum.
- Distribution- and density-based clustering: e.g., DBSCAN.
Clustering

Clustering (Unsupervised Learning)

- Hierarchical: agglomerative or divisive.
- $k$-means: NP-hard, so algorithms find a local minimum.
- Distribution- and density-based clustering: e.g., DBSCAN.
- Fuzzy clustering: membership is not binary.
Clustering ... and Classification

Clustering (Unsupervised Learning)
- Heirarchical: agglomerative or divisive.
- $k$-means: NP-hard, so algorithms find a local minimum.
- Distribution- and density-based clustering: e.g., DBSCAN.
- Fuzzy clustering: membership is not binary.

Classification (Supervised Learning)

input data (training sample): $D = \{(X_i, Y_i)\}_{i=1}^n$

$k$-nn clustering: for new $X$, we predict $Y$ by majority vote of the $k$ nearest neighbors of the covariates (features) in $D$. 
Curate a list of topological descriptors. For each, we are looking for:

- Name of descriptor.
- List of distances that can be used between descriptors.
- Short explanation (very short).
- Reference to where first used, or a good use of it.
- Pros: What is it good for?
- Cons: Where / when is it insufficient?

https://github.com/compTAG/ima-multid