Classification of Stochastic Processes with Topological Data Analysis

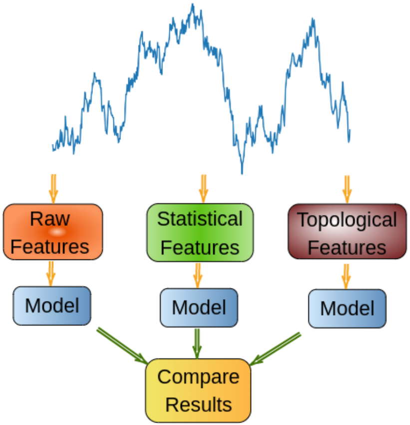
İsmal Güzel & Atabey Kaygun

Mathematics Engineering - İTÜ

May 11, 2022

Outline

1. Stochastic Process
2. Topological Data Analysis
3. Feature Engineering
4. Results
   - Balanced and Unbalanced



Stochastic Processes

Wiener Process

- For $0 \leq s < t < u < v \leq T$, $X_t - X_s$ and $X_v - X_u$ are independent.
- For $0 \leq s < t \leq T$,

$$X_t - X_s \sim \sqrt{t-s}N(0,1)$$

Cauchy Process

- For $0 \leq s < t < u < v \leq T$, $X_t - X_s$ and $X_v - X_u$ are independent.
- For $0 \leq s < t \leq T$,

$$X_t - X_s \sim Cauchy(t-s; 0, 1)$$

Cauchy process is a Brownian motion with a Levy subordinator.

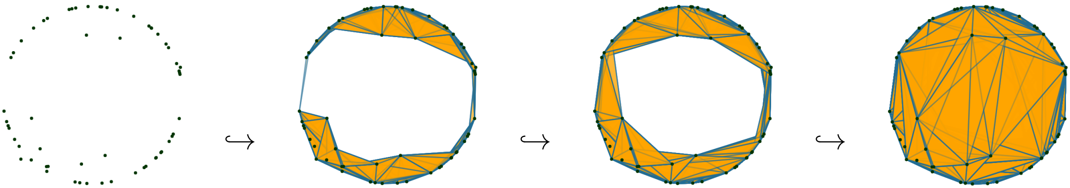
Can we distinguish these processes by using statistical or topological features?

[Applebaum, 2009]

Topological Structure

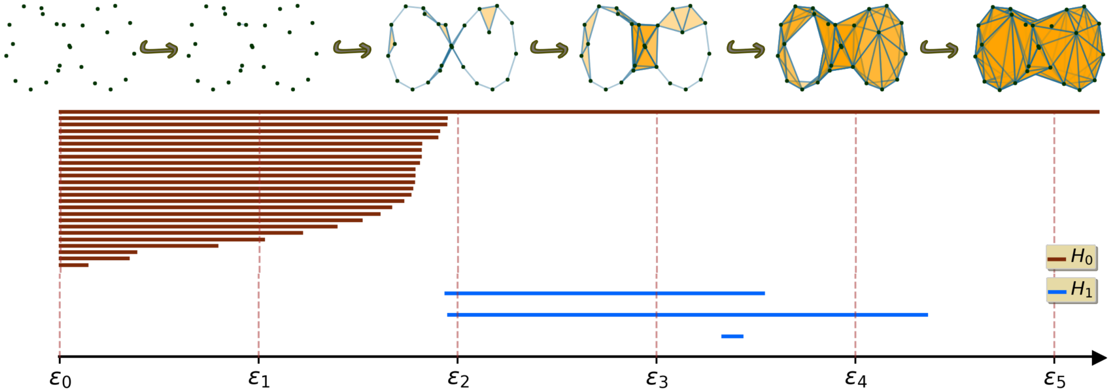
Given a point cloud X, the Vietoris-Rips is defined to be the simplicial complex whose simplices are built on vertices that are at most ε apart,

$$R_\varepsilon(X) = \{\sigma \subset X \mid d(x,y) \leq \varepsilon, \text{ for all } x, y \in \sigma\}.$$



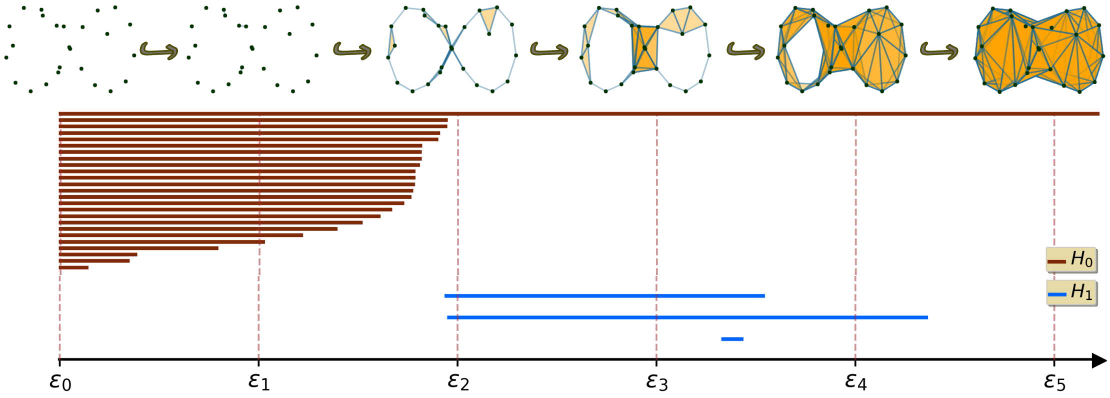
$\beta_0 = 50$	$\beta_0 = 1$	$\beta_0 = 1$	$\beta_0 = 1$
$\beta_1 = 0$	$\beta_1 = 1$	$\beta_1 = 1$	$\beta_1 = 0$

Persistent Homology[1] and Persistence Barcodes[2]

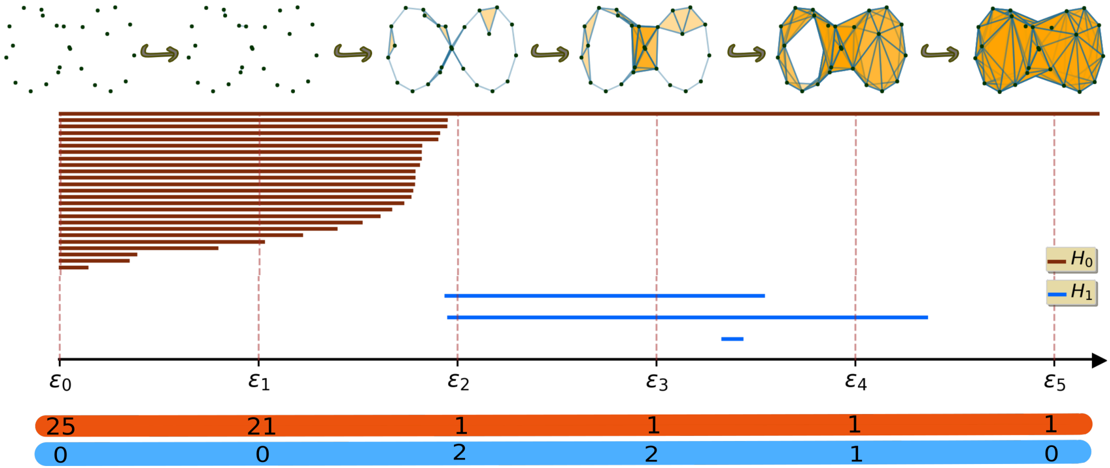


[1] [Carlsson, 2009]
[2] [Ghrist, 2008]

Persistent Homology and Persistence Barcodes

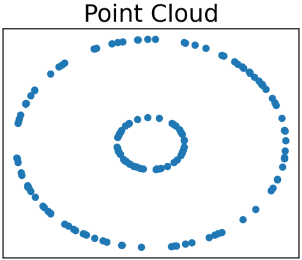


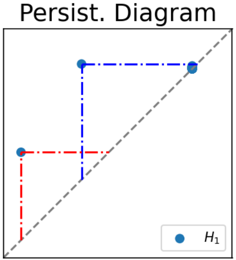
Persistence Barcodes and Betti Curves



Persistence Landscapes

- Diagram $D = \{(a_i, b_i)\}_{i \in I}$ and for $a < b$
- $f_{(a,b)}(t) = \max(0, \min(a + t, b - t))$
- $\lambda_k(t) = \text{kmax}\left\{f_{(a_i,b_i)}(t)\right\}_{i \in I}$

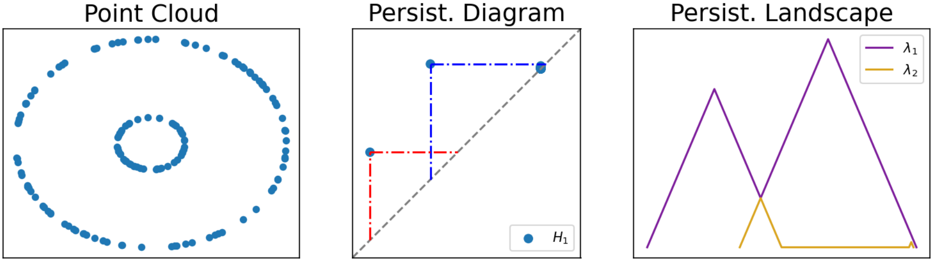




[Bubenik, 2020]

Persistence Landscapes

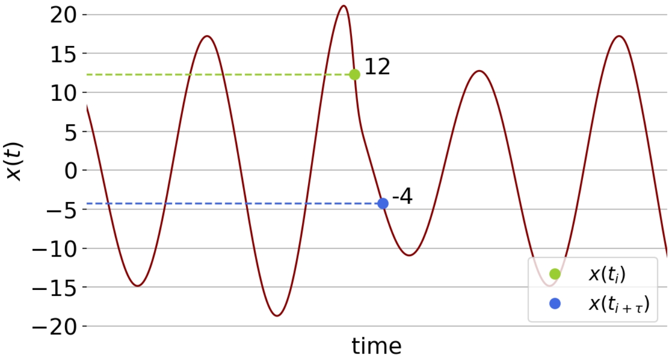
- Diagram $D = \{(a_i, b_i)\}_{i \in I}$ and for $a < b$
- $f_{(a,b)}(t) = \max(0, \min(a + t, b - t))$
- $\lambda_k(t) = \mathrm{kmax}\left\{f_{(a_i,b_i)}(t)\right\}_{i \in I}$



[Bubenik, 2020]

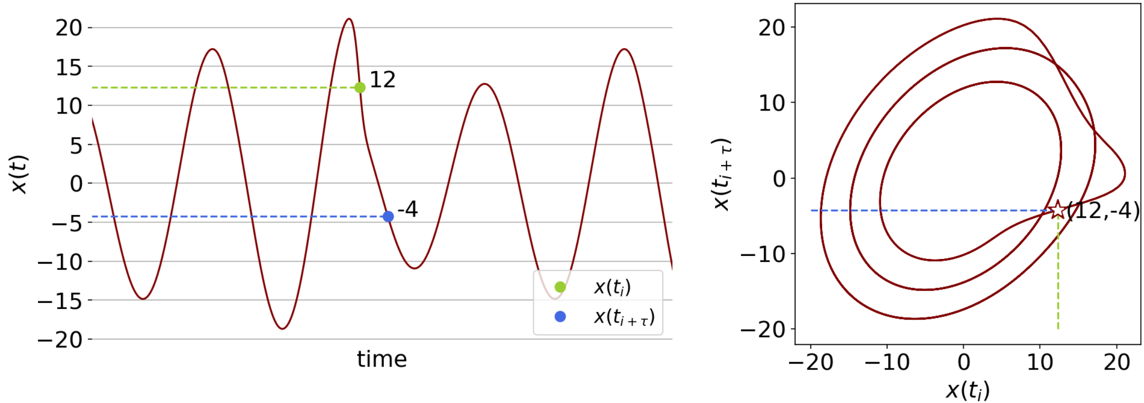
Takens' Delay Embedding

- $T = (x_1, x_2, x_3, \cdots, x_n) \quad \Longrightarrow \quad PC = \{v_i\}_i$, where $v_i = \{x_i, x_{i+\tau}, \ldots, x_{i+(d-1)\tau}\}$
- Parameters: $\tau = 3$ and $d = 2$
- $PC = \{(x_1, x_4), (x_2, x_5), \cdots, (x_{n-3}, x_n)\}$

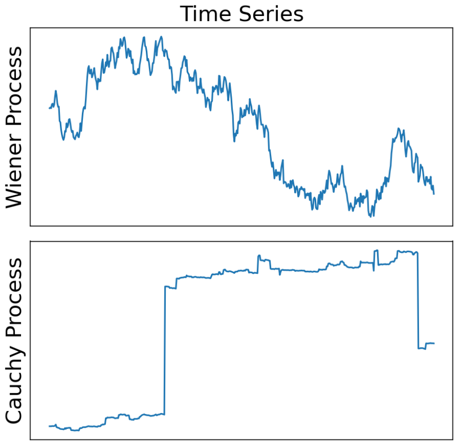


Takens' Delay Embedding

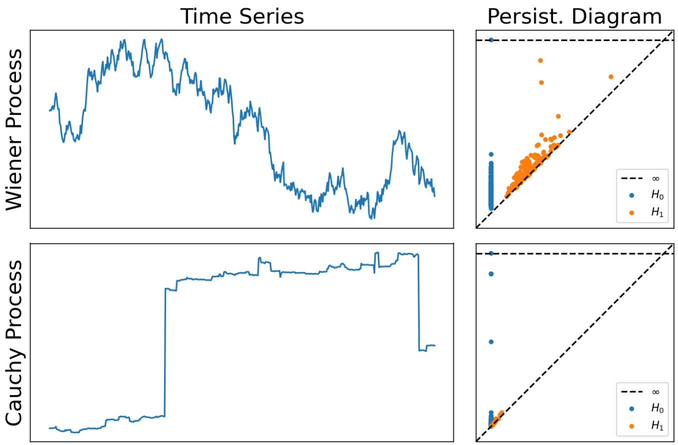
- $T = (x_1, x_2, x_3, \cdots, x_n) \quad \Longrightarrow \quad PC = \{v_i\}_i$, where $v_i = \{x_i, x_{i+\tau}, \ldots, x_{i+(d-1)\tau}\}$
- Parameters: $\tau = 3$ and $d = 2$
- $PC = \{(x_1, x_4), (x_2, x_5), \cdots, (x_{n-3}, x_n)\}$



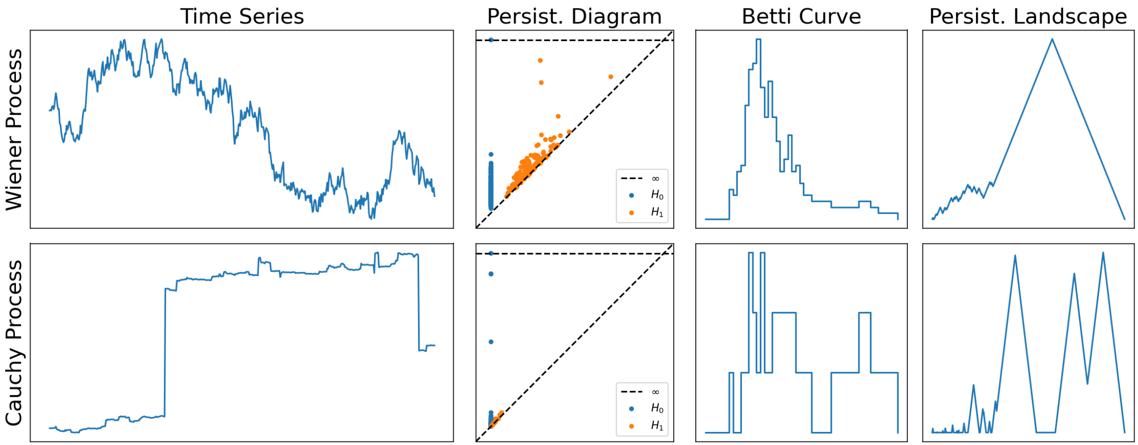
Topological Engineered Features



Topological Engineered Features



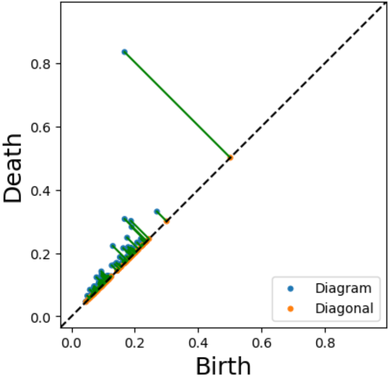
Topological Engineered Features



Topological Engineered Features

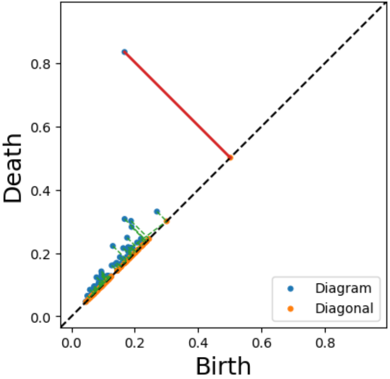
Wasserstein Distance

$$W_1(D, D_\emptyset) = \sum_{i \in I} \frac{d_i - b_i}{2}$$



Bottleneck Distance

$$d_B(D, D_\emptyset) = \sup_{i \in I} \frac{d_i - b_i}{2}$$



Topological Engineered Features

From the persistence diagram $D = \{(b_i, d_i)\}_{i\in\mathbb{I}}$ with $\ell_i = d_i - b_i$ and $L_D = \sum_{i\in\mathbb{I}} \ell_i$.

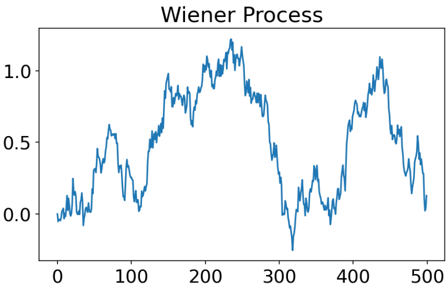
Adcock-Carlsson Coordinates

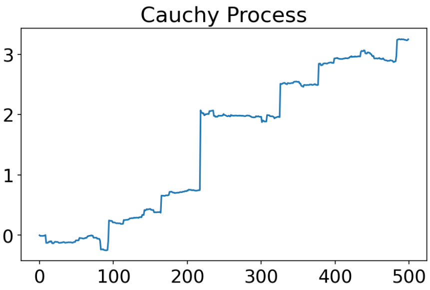
- $f_1(D) = \sum_{i\in\mathbb{I}} b_i \ell_i$
- $f_2(D) = \sum_{i\in\mathbb{I}} (d_{max} - d_i)\ell_i$
- $f_3(D) = \sum_{i\in\mathbb{I}} b_i^2 \ell_i^4$
- $f_4(D) = \sum_{i\in\mathbb{I}} (d_{max} - d_i)^2 \ell_i^4$

Persistence Entropy

$$E(D) = -\sum_i \frac{\ell_i}{L_D} \log\left(\frac{\ell_i}{L_D}\right)$$

Experiments





Simulation

- Generate 1000 time series with same length of 500
- Label to *Wiener* and *Cauchy*

Feature Engineering

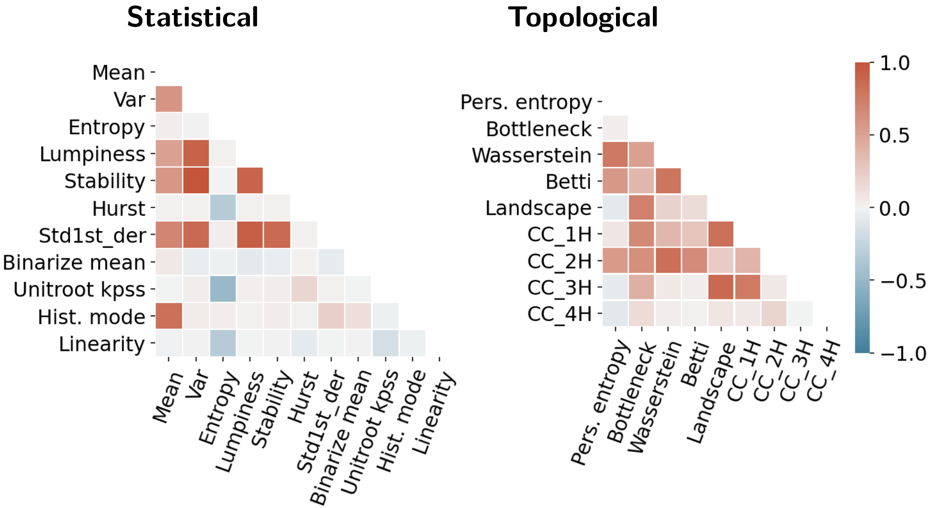
- Raw features
- Statistical features

1. Mean
2. Variance
3. Entropy
4. Lumpiness
5. Stability
6. Hurst
7. Std 1st-der
8. Linearity
9. Binarize mean
10. Unitroot KPSS
11. Histogram mode

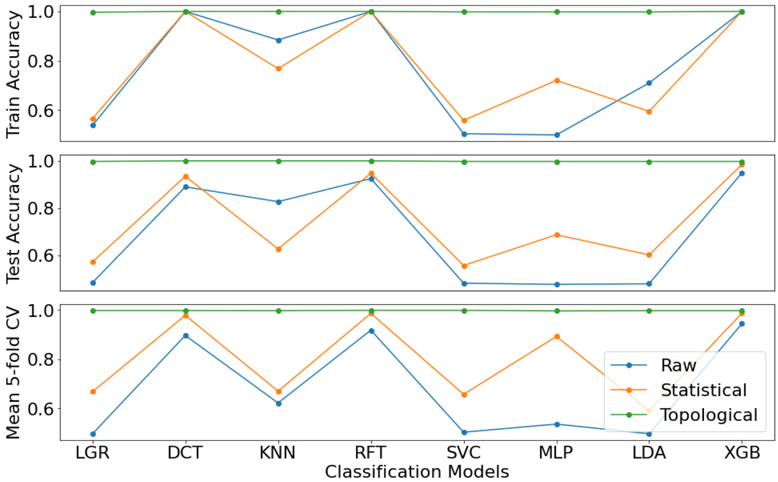
- Topological features

1. Bottlenecek
2. Wasserstein
3. Persistence entropy
4. f_1
5. f_2
6. f_3
7. f_4
8. L_1 Norm of Betti curve
9. L_1 Norm of Persistence landscapes

Correlation between features



Classification algorithms on balanced dataset

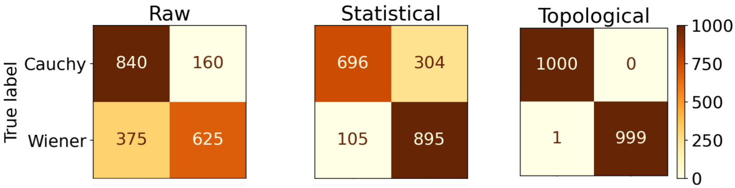


Classification Models:

- LGR: Logistic Regression
- DCT: Decision Tree
- KNN: k-Nearest Neighbor
- RFT: Ranfom Forest
- SVC: Support Vector
- MLP: Multi Layer
- LDA: Linear Discriminant
- XGB: XGBoost

Confusion matrices for KNN

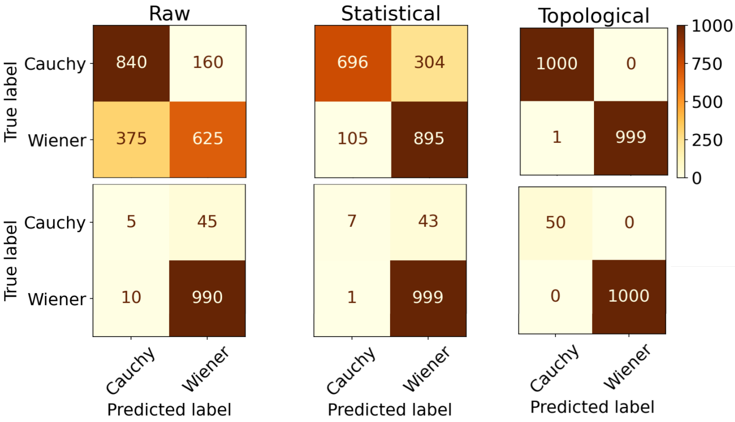
Balanced Dataset



Confusion matrices for KNN

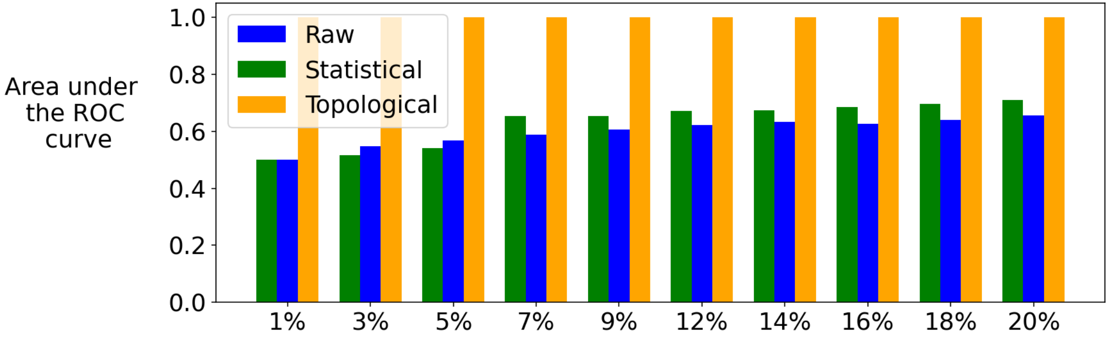
Balanced Dataset

Unbalanced Dataset



Different rates on the unbalanced dataset

Majority class (Wiener) has a sample size : 1000
Minority class (Cauchy) is varied between 1% and 20% of the majority class



What about the computation cost?

We thank Turkish National Center for High Performance Computing (UHeM),
Cluster Features:
Intel Xeon E5-2680 CPU (28 Cores) with 128GB RAM

We test on:

- 50 Cauchy processes
- with randomly changing length between 500 and 1500
- for each experiments run 7 times

Features		**Mean**	**Std**
Topological	Parallel	45.9 s	321 ms
	Serial	111 s	1900 ms
Statistical	Parallel	1.12 s	40.3 ms
	Serial	2.10 s	3.16 ms

Future Work

1. Theoretical explanation of Levy subordinator.
2. Compare with other features such as discrete Fourier transforms, power spectral densities, etc.
3. Apply real world problems.

References

Applebaum, D. (2009).
Lévy processes and stochastic calculus.
Cambridge university press.

Bubenik, P. (2020).
The persistence landscape and some of its properties.
In *Topological data analysis. Proceedings of the Abel symposium 2018, Geiranger, Norway, June 4–8, 2018*, pages 97–117. Cham: Springer.

Carlsson, G. (2009).
Topology and data.
Bulletin of the American Mathematical Society, 46(2):255–308.

Ghrist, R. (2008).
Barcodes: the persistent topology of data.
Bulletin of the American Mathematical Society, 45(1):61–75.

Thank You!

iguzel@itu.edu.tr