# MATH 5473 Project: Exploring Time-Series Data: Visualization, Clustering, and Topological Analysis with an SNP500 Case Study

Anastasiia KAZOVSKAIA[1], Songjie XIE[2] and Meiying ZHANG[2] {akazovskaia, sxieat, mzhangcp}@connect.ust.hk
[1]Department of Mathematics, HKUST, [2] Department of Electronic and Computer Engineering, HKUST
Contribution: Anastasiia KAZOVSKAIA: time series analysis and topological analysis exploration, Songjie XIE: visualization and clustering, Meiying ZHANG: data analysis and poster

## 1. Introduction

In the modern financial landscape, the analysis of time-series data plays a crucial role in understanding market trends, predicting future movements, and making informed investment decisions. Time-series data, characterized by its sequential nature and inherent variability, presents unique challenges in terms of analysis and interpretation. In this study, we aimed to explore time-series data, using a case study of the S&P 500, a widely recognized stock market index. Our objective was to apply a combination of data visualization techniques, clustering methods, and topological data analysis (TDA) to extract meaningful insights.

## 2. Dataset

SNP'500 is a stock market index that tracks the performance of 500 large companies in the United States. The finance dataset SNP'500 contains 1258-by-452 matrix, which represents the closed price of stocks from 452 American company in 1258 consecutive workdays.

## 3. Data Preprocessing

we employed the following preprocessing techniques for time-series data, S&P 500 :

1. *Normalization*: It includes min-max and z-score normalization.
2. *Differentiation*: first-order and second-order differentiations.
3. *Log-return*: We calculated the log percentage change of the time series, which is a commonly used method in finance.
4. *Extracted feature*: the features of time series are interpretable by statistical analysis.
5. Smoothing: It help us better investigate patterns and trends in time series, which will benefits the TDA .

## 4. Visualization

The data visualization is achieved by employing dimensionality reduction, including PCA, kernel PCA, UMAP, t-SNE, MDS, and ISOMAP, on the row data, extracted features, and log return.

➢ Visualization Experimental Results

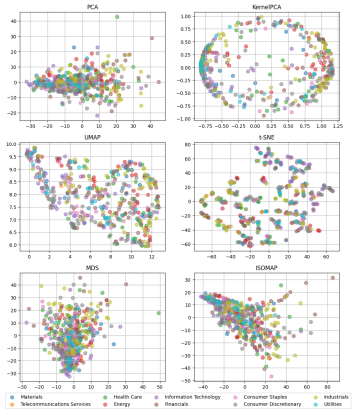


Fig 1. Results of dimensionality reduction on row data

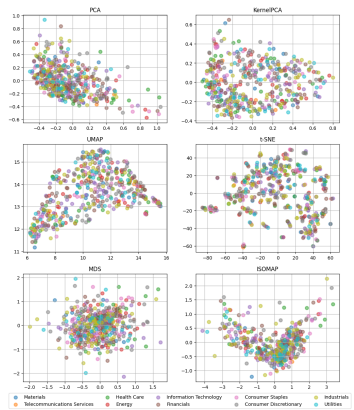


Fig 2. Results of dimensionality reduction on features of each time series

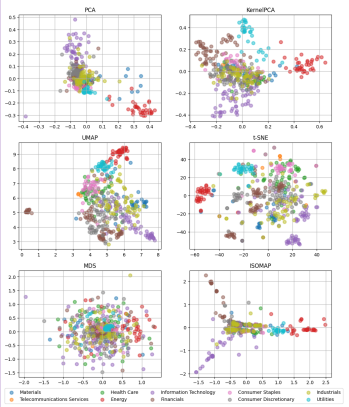


Fig 3. Results of dimensionality reduction on log return of each time series

➢ Analysis

Upon comparison of the visualization results, we observed that visualizations based on log returns provided more meaningful insights and clearer patterns. Because the resulting dataset mitigates the effects of differing scales among various stocks by converting price data into log returns. This normalization helps ensure that the dimensionality reduction techniques focus on the relative changes in price rather than absolute differences, leading to more consistent visualizations.

## 5. Clustering

For clustering time-series data, we employ standard clustering methods such as k-means and k-shape, as well as the high-dimensional approach known as sparse subspace clustering (SSC). Additionally, we utilize different preprocessing techniques mainly categorized into time-domain representation and feature-based presentations prior to applying clustering.

➢ Clustering Experimental Results

CLUSTERING PERFORMANCE ON THE FINANCIAL DATASET. NUMBERS IN BOLD AND UNDERLINED OUTLINE THE BEST AND RUNNER-UP RESULTS, RESPECTIVELY.

| Data | Preprocessing | Method | ACC(%) | ARI(%) | $v$-score | NMI(%) | H-score |
|---|---|---|---|---|---|---|---|
| time series | norm* | $k$-means | 19.69 | 0.89 | 8.46 | 8.46 | 8.17 |
| | norm* | $k$-shape | 20.13 | 1.47 | 8.76 | 8.76 | 8.50 |
| | $diff_1$ | SSC | **58.63** | **35.88** | **54.53** | **54.53** | **53.86** |
| | $diff_2$ | SSC | 53.32 | 30.65 | 50.99 | 50.99 | 50.23 |
| | log return + PCA | $k$-means | 42.48 | 19.35 | 39.36 | 39.36 | 36.31 |
| | log return + KernelPCA | $k$-means | 46.46 | 22.55 | 42.14 | 42.14 | 42.17 |
| | log return + UMAP | $k$-means | 45.58 | 28.96 | 48.87 | 48.87 | 49.36 |
| | log return + $t$-SNE | $k$-means | 44.03 | 26.88 | 45.65 | 45.65 | 45.85 |
| | log return + MDS | $k$-means | 26.99 | 7.62 | 17.56 | 17.56 | 17.11 |
| | log return + ISOMAP | $k$-means | 40.27 | 19.12 | 38.32 | 38.32 | 36.87 |
| feature-based | norm | $k$-means | 17.92 | -0.30 | 3.42 | 3.42 | 3.02 |
| | log return + PCA | $k$-means | 17.69 | -0.34 | 3.38 | 3.38 | 3.01 |
| | log return + KernelPCA | $k$-means | 16.81 | -0.33 | 4.32 | 4.32 | 3.84 |
| | log return + UMAP | $k$-means | 15.27 | -0.20 | 3.76 | 3.76 | 3.84 |
| | log return + $t$-SNE | $k$-means | 16.15 | -0.07 | 4.04 | 4.04 | 4.13 |
| | log return + MDS | $k$-means | 17.26 | -0.37 | 3.56 | 3.56 | 3.18 |
| | log return + ISOMAP | $k$-means | 17.92 | -0.23 | 3.72 | 3.72 | 3.27 |

Table 1. Results of Clustering

➢ Analysis

Upon comparing the results, it is observed that combining log return representations with dimensionality reduction methods yields favorable performance. Surprisingly, the SSC technique with first-order differentiation achieves the highest level of performance. Conversely, feature-based and high-dimensional representations demonstrate poor performance.

## 8. References

[1] Harold Hotelling. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417, 1933.
[2] J. B. Kruskal. Multidimensional scaling by optimizing goodness of ift to a nonmetric hypothesis. Psychometrika, 29(1):1–27, Mar 1964.
[3] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.
[4] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. science, 290(5500):2319–2323, 2000.
[5] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

## 6. Topological analysis exploration

The time series data is generally challenging to visualize and cluster, we explore how to use TDA to analyze the time series data.

➢ TDA Example

Fig 4. Original time series

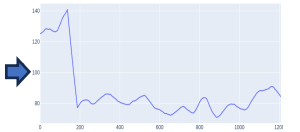

Fig 5. Smoothed time series

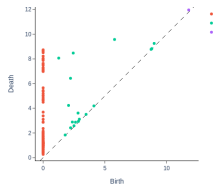


Fig 7. Persistence diagram from Vietoris-Rips filtration

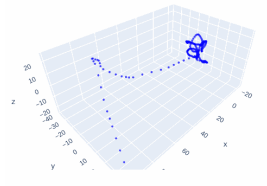

Fig 6. Topological representation of time series

➢ Analysis

1. Smooth time series to reduce noise and emphasize key trends and patterns.
2. Convert time series into time delay embeddings.
3. Use PCA to visualize the embeddings.
4. Obtain persistence diagrams to study topological properties of point clouds.

## 7. Conclusion

Time-series data presents ongoing challenges, and our project has revealed the absence of a universally effective method. Whether using feature-based analysis, topological methods, or high-dimensional approaches, the selection of techniques must be tailored to each specific case. A case-by-case approach is essential when working with time-series data.