Comprehensive Analysis of Wave Data Using Principal Component Analysis

Hung Tran

May 8, 2024

1 Introduction

This report presents an in-depth analysis of a dataset that includes measurements from 5005 wave experiments, each characterized by 40 distinct features. My objective is to apply Principal Component Analysis (PCA) to decrease the number of dimensions in the dataset, capture important information, and distinguish various types of waves.

2 Analysis

2.1 Singular Value Decomposition & Component Analysis

The data matrix B was analyzed using Singular Value Decomposition (SVD). The singular values were plotted on a logarithmic scale to identify the decay rate and determine the number of principal components that capture the most variance.

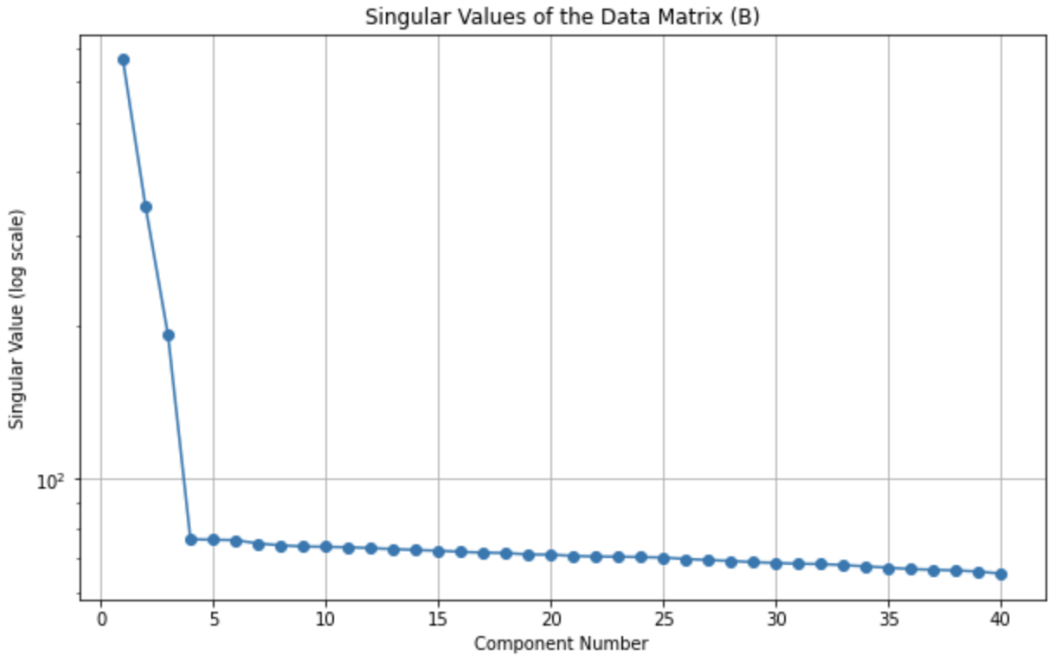


Figure 1: Singular values of the data matrix plotted on a logarithmic scale.

When plotted on a logarithmic scale (Figure 1), these values showed a sharp decline, suggesting that only the first few singular values are significant, while the others contribute much less to the data's variance. It implies the presence of a lower-dimensional structure in the high-dimensional dataset.

2.2 Deciding the Dimension q for PCA

To decide on the dimension q for reducing the dataset, we examine where additional singular values significantly contribute to the explained variance. A common approach involves selecting q such that a substantial proportion (90%) of the variance is captured. This method ensures that the reduced dimensionality retains the most significant information from the original dataset.

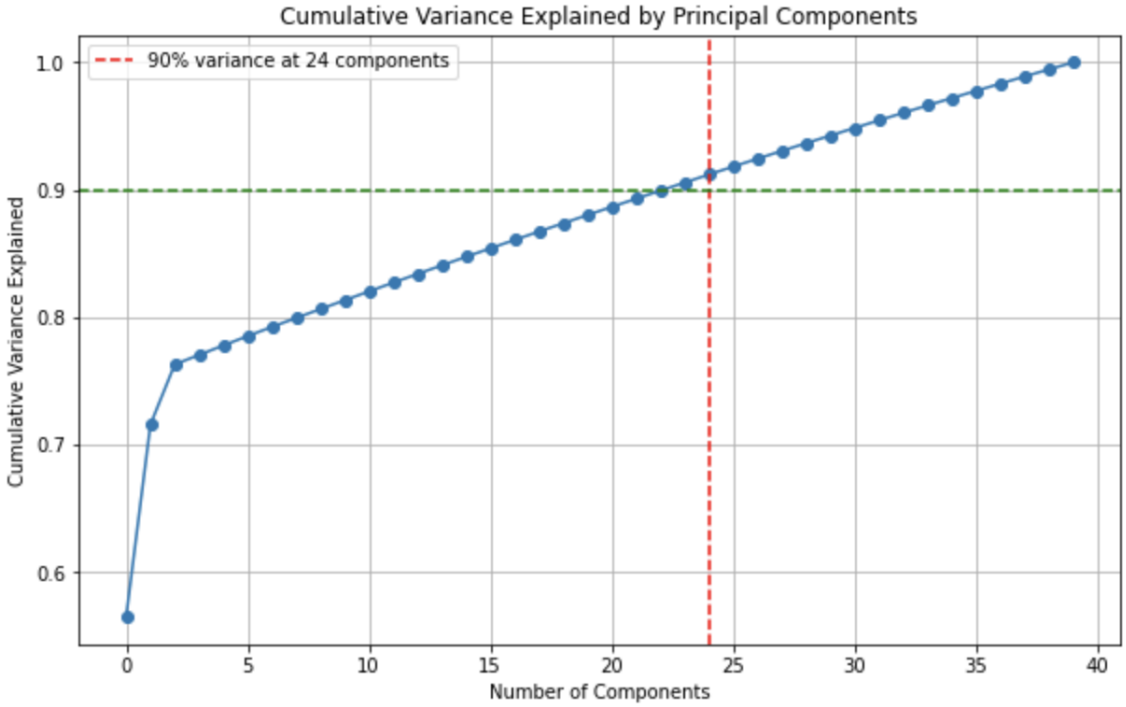


Figure 2: Cumulative variance explained by the principal components.

The plot shows a point where adding more components does not substantially increase the explained variance (see Figure 2). 24 principal components are enough to account for 90% of the dataset's variance. This finding supports the idea of reducing the dimensionality to $q = 24$, effectively retaining most of the essential information while making the dataset much simpler to manage. Such a reduction not only makes data processing and analysis more efficient but also simplifies the visualization and understanding of the key patterns in the data.

2.3 Projection and Visualization

The dataset was then projected onto the first 24 principal components. Following this, visualizations were created to further analyze the structure of the data:

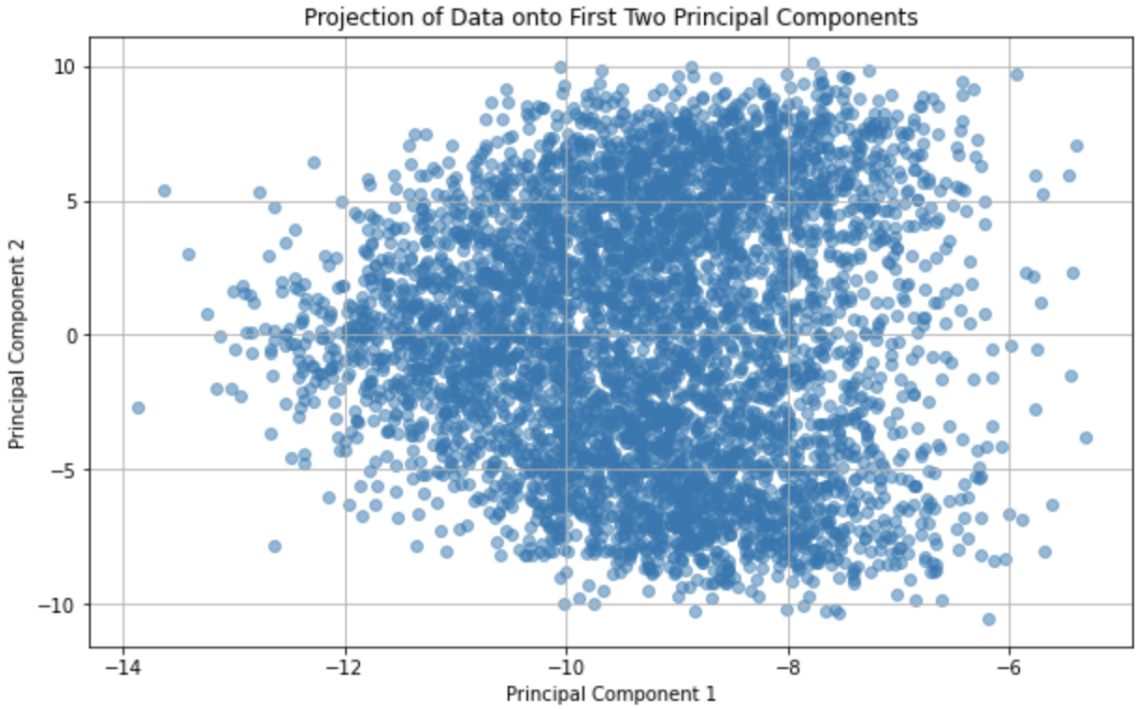


Figure 3: 2D PCA projection of the data onto the first two principal components

The scatter plot of the data projected onto the first two principal components reveals some clustering, which suggests that the dataset might contain distinct groups or types of waves.

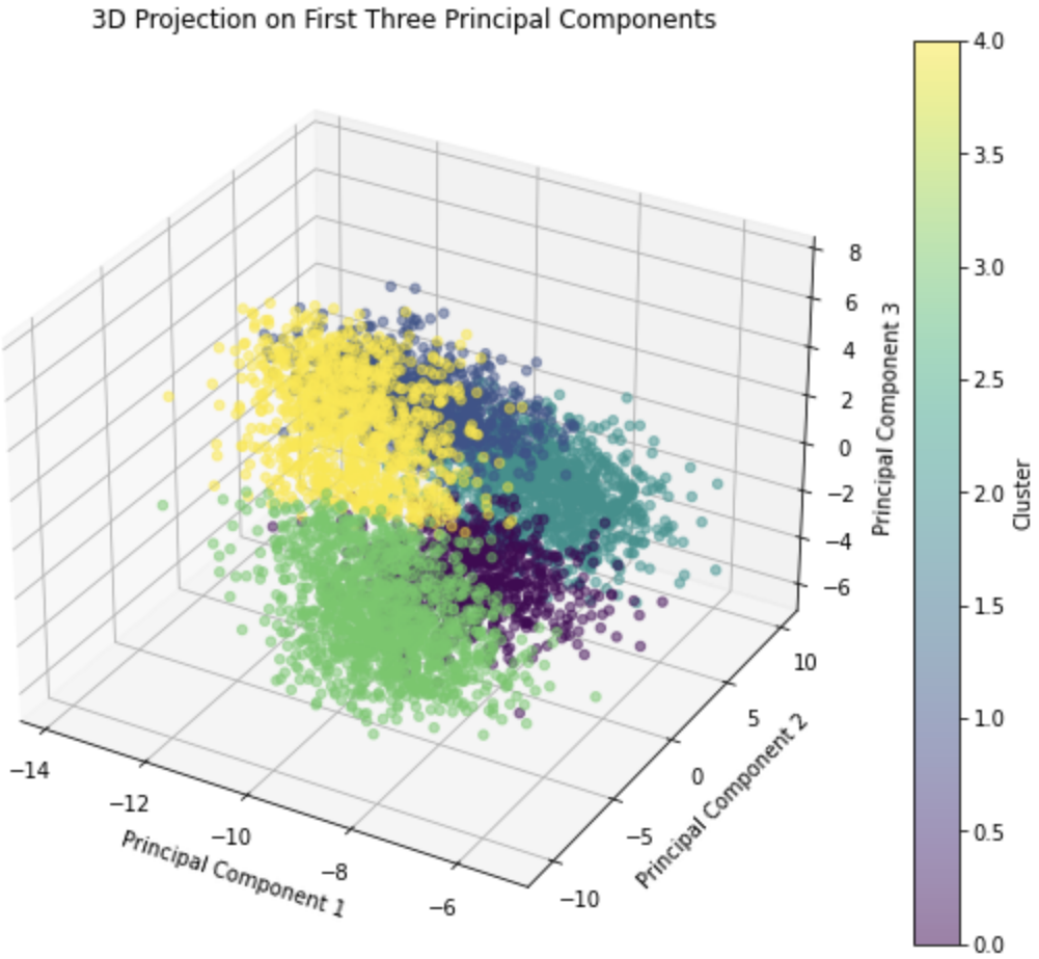
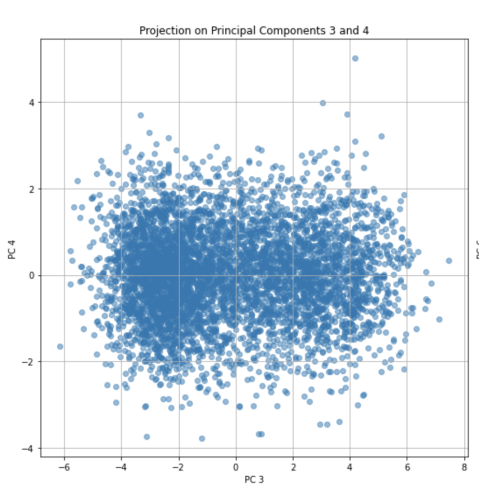
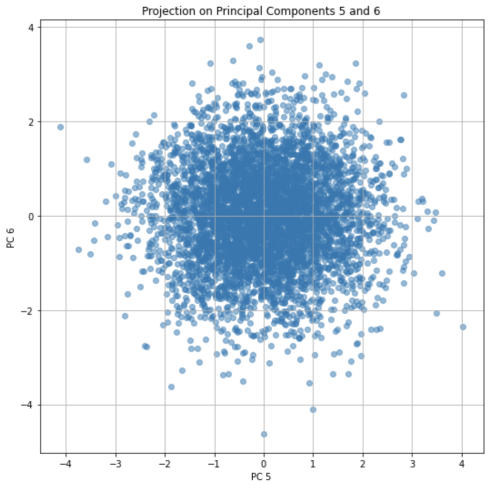


Figure 4: 3D PCA projection of the data onto the first three principal components

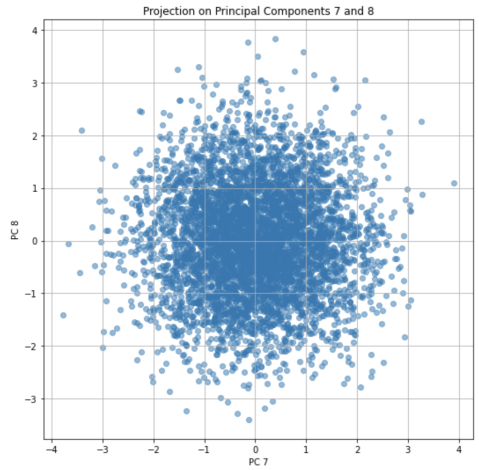
The plot ensures the presence of distinct wave types and highlights the complexity of their interactions.



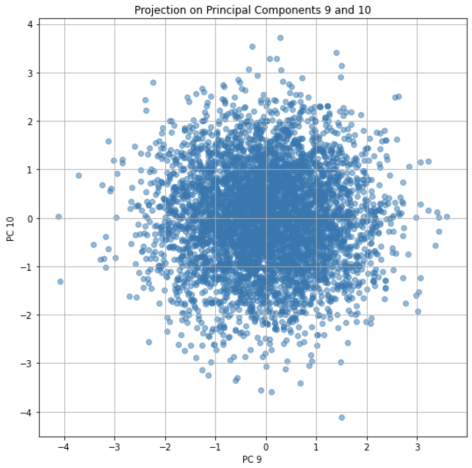
(a) Projection on PC3 and PC4



(b) Projection on PC5 and PC6



(c) Projection on PC7 and PC8



(d) Projection on PC9 and PC10

Figure 5: Projections on different combinations of principal components

The scatter plots for different combinations of principal components (PCs 3 and 4, PCs 5 and 6, PCs 7 and 8, PCs 9 and 10) show varied patterns, with some displaying more clear clustering than others. These visualizations suggest that different characteristics of the data may be captured by different principal components.

2.4 Clustering Analysis

K-means clustering was applied with 5 clusters to the data reduced to 24 dimensions, and visualizing the results on the first two principal components reveals distinct clusters. These clusters could represent different types of wave patterns as initially hypothesized.

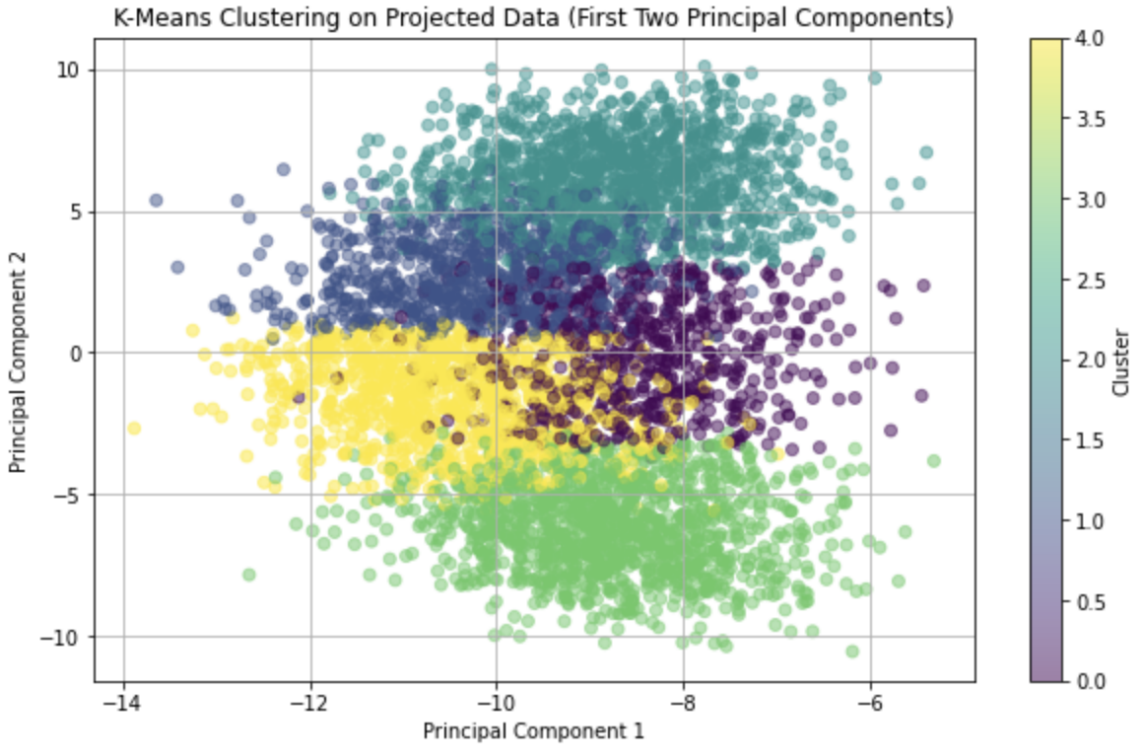


Figure 6: Clustering results on the reduced data.

The clustering highlights the distinct groups formed in the lower-dimensional space. It appears reasonably well defined, indicating effective dimensionality reduction and clustering.

3 Conclusion

The application of PCA to this dataset effectively lowered the number of dimensions while preserving more than 90% of the total variance, offering a clearer and more practical perspective on the data's structure. The resulting projections and clustering analysis highlight distinct groupings within the data, indicating different types of waves. This insight could be extremely useful for more detailed studies in wave analysis.

4 Bonus Analysis: Fashion-MNIST Dataset

4.1 Singular Value Decomposition and Component Analysis

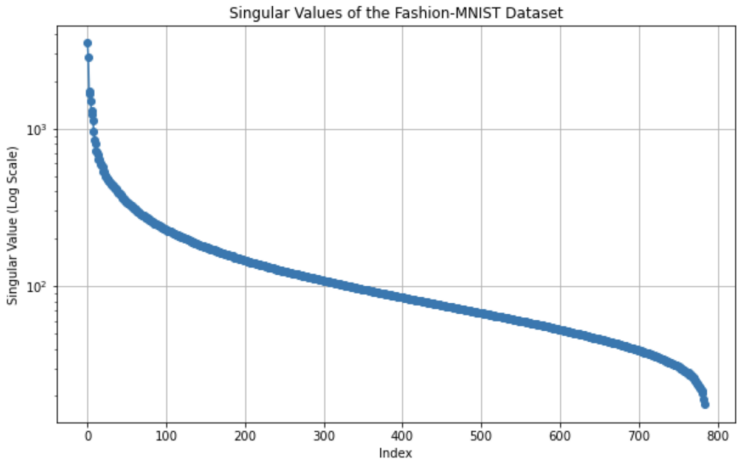


Figure 7: Singular values of the Fashion-MNIST dataset plotted on a logarithmic scale

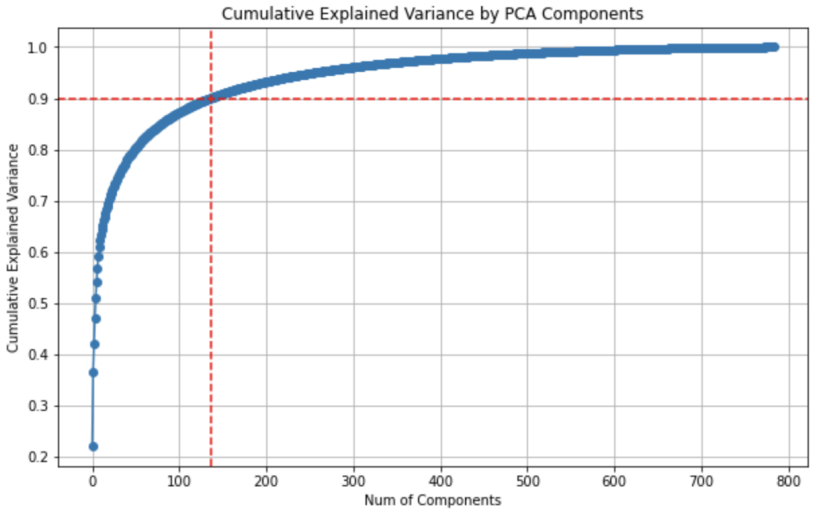


Figure 8: Cumulative Explained Variance by PCA Components, $q = 137$