Rachel Yang
Laura Lall

Introduction to Machine Learning
Practical 1 - Dimensionality Reduction Using PCA

Introduction

In this assignment we implement a Principal Component Analysis (PCA) to reduce the dimensionality of the given dataset ‘COIL20.mat’, which contains 1440 images of size 32x32 pixels, flattened to a vector of 1440 dimensions. By applying PCA, we aim to retain the most important features of the data while reducing its dimensionality. This report documents the implementation of PCA, the experimental results, and a discussion of the findings.

Methods

1. **Standardization**
   The dataset X is standardized by subtracting the mean (μ) and dividing by the standard deviation (σ) for each feature:

$$Z = \frac{X - \mu}{\sigma}$$

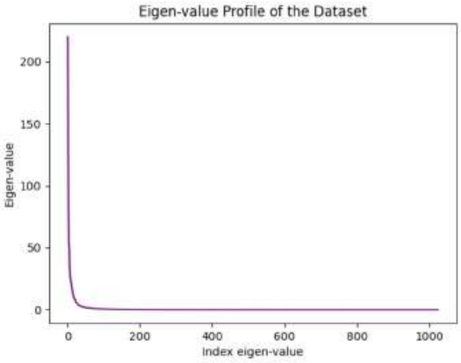
2. **Covariance matrix**
   The covariance matrix of the standardized data Z is computed to understand the relationship between features.

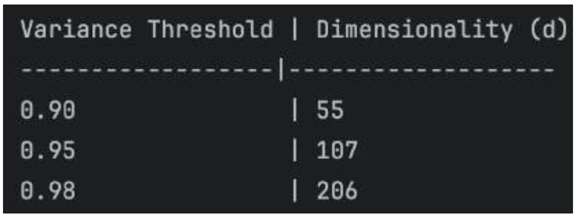
3. **Eigenvalue decomposition**
   The eigenvalues and eigenvectors of the covariance are computed. The eigenvectors represent the principal components, and the eigenvalues indicate the amount of variance captured by each component.

4. **Dimensionality reduction**
   The top d eigenvectors are selected based on the largest eigenvalues. The data are then projected onto these components to obtain the redacted dataset Z_d.

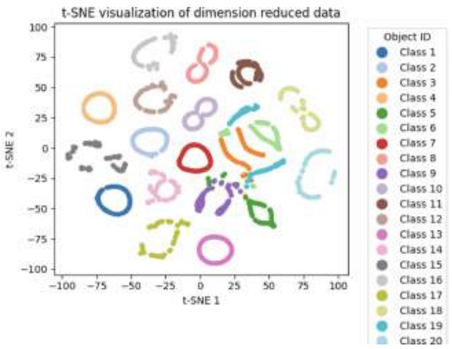
$$Z_d = U_d^T Z$$

Experimental results





```
Variance Threshold | Dimensionality (d)
------------------|-------------------
0.90              | 55
0.95              | 107
0.98              | 206
```



Discussion

Eigenvalue profile analysis:

The eigenvalue profile reveals that the eigenvalues decline steeply, the higher the index gets the slower the decline becomes. This indicates that only a small number of principal components capture the most variance in the dataset, while remaining ones contribute minimally. The rapid drop shows that dimensionality reduction using PCA is effective, since a big portion of the dataset's variance can be retained using relatively few components.

Dimensionality Selection for Variance Retention

It is important to select an appropriate number of dimensions, since we need to avoid significant memory loss or creating unnecessary computational complexity. Our analysis shows that:

- 55 dimensions are required to retain 90% of the total variance.
- 107 dimensions are needed to retain 95% of the variance.
- 206 dimensions are necessary to retain 98% of the variance.

This result aligns with the eigenvalue profile, confirming that the variance is concentrated in the first few principal components. This further supports the effectiveness of PCA in reducing dimensionality while preserving essential image.

t-SNE Visualization of Reduced Data

The t-SNE visualization of the dataset (with d =40) shows clear clusters corresponding to different objects. This indicates that the redacted representation preserves the underlying structure of the data, as distinct objects are well-separated in the 2D feature space. The successful clustering demonstrates that PCA retains essential structural information, even after significant dimensionality reduction.

Conclusion

In conclusion, based on our experimental results we can confirm that PCA reduced dimensionality effectively, while retaining essential variance. For this assignment Rachel wrote most of the code, while Laura added to it, fixed mistakes, commented on the code and made it compatible with the requirements in Themis. The report was written mostly by Laura, Rachel added to it and fixed any mistakes. In the upcoming assignments we will make sure to switch roles and also try to work on both things together more.