AMATH 482/582: HOMEWORK 2

XINYUE LYU

Amazing Department, University of Washington, Seattle, WA
xlyu6@uw.edu

Abstract. This report presents the analysis of human movement data using Principal Component Analysis (PCA) for dimensionality reduction and classification. Motion data from walking, running, and jumping activities were processed and projected in a lower-dimensional space. The results demonstrate the effectiveness of PCA in classification, providing insights into human activity recognition. The findings highlight the role of machine learning in analyzing motion data efficiently.

1. Introduction and Overview

Human activity recognition is a crucial problem in machine learning, with applications in healthcare, sports analytics, and human-computer interaction. This assignment explores the use of Principal Component Analysis (PCA) for dimensionality reduction and classification of movement data. In this assignment, I analyze three movement types: walking, running, and jumping. I first investigate how many PCA spatial modes are needed to keep to approximate train data up to 70% 80% , 90% , 95% in Frobenius norm. By projecting the data into a lower-dimensional space using PCA, I aim to design an algorithm that will be able to recognize which movement OptimuS-VD robot is performing in real-time.

2. Theoretical Background

Singular value decomposition(SVD) is a fundamental technique in linear algebra and data analysis, widely used for dimensionality reduction and data compression. Given a matrix A of shape $m \times n$, SVD decomposes it into three matrices:

$$A = U\Sigma V^T \tag{1}$$

where:

- U $(m \times m)$ contains the left singular vectors (eigenvectors of AA^T), a rotation,
- Σ $(m \times n)$ is a diagonal matrix of singular values, a stretching, representing the importance of each mode and order the value from high to low,
- V^T $(n \times n)$ contains the right singular vectors (eigenvectors of A^TA), a rotation.

PCA finds a set of orthogonal basis vectors, called principal components, which capture the directions of maximum variance in the data. Given a dataset represented as a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ (where rows represent observations and columns represent features), PCA is performed by computing the covariance matrix:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{X}.$$

Generally, people assume that data measurements with a high variance capture more important

Date: February 13, 2025.

features. PCA can be computed using SVD 1. If we take the SVD of the mean-centered data matrix. Then the principal components correspond to the right singular vectors, and the singular values in $\mathbf{\Sigma}$ determine the amount of variance captured by each component.

By selecting only the first k principal components, truncating the PCA modes, we can reduce the dimensions of the data while preserving most of its variance. This allows us to visualize high-dimensional structures in lower-dimensional spaces.

In this assignment, I apply PCA to movement data (walking, running, and jumping) in 100 time steps to extract dominant motion patterns. By truncating the PCA modes, I analyze how different levels of dimensionality affect the movement trajectories and the accuracy of the model.

3. Algorithm Implementation and Development

In this assignment, I implemented Principal Component Analysis (PCA) for dimensionality reduction from the movement data. The PCA algorithm was applied using the `scikit-learn` package, which provides efficient tools for data preprocessing and decomposition. For data handling and manipulation, I utilized `numpy` for structured data storage and efficient numerical operations like computing cumulative sum of energy. Visualization of results was conducted using `matplotlib`, allowing for clear representations of PCA projections.

4. Computational Results

To determine the number of PCA modes required to retain different percentages of the dataset's total energy(Frobenius norm), I computed the cumulative sum of normalized singular values. The results indicate that a small number of principal components capture most of the variance. Specifically, retaining 70% of the energy requires only a few modes, whereas achieving 95% energy retention demands a greater number of modes.

Energy	Number of PCA modes
70 percent	2
80 percent	3
90 percent	5
95 percent	7

Table 1. This table shows how many dimentions(modes) in PCA space are needed to approximate the train data up to four different thresholds.

The Figure 1 demonstrates that a relatively small number of PCA modes, fewer than 10 modes, can capture a significant portion of the variance in the joint movement data, enabling effective dimensionality reduction for motion recognition tasks.

Then, I set the number of PCA modes be 2 and 3, and plot the projected train data of the movements in PCA space as 2 and 3 dimensional trajectories. Figure 2 revealed distinct clusters corresponding to the trajectory of the movement over time. The red dots are running movement. They located in the area with PC2 values around 500 to 1000 and PC1 values around -1500 to 1500. The black dots are walking movement. They located in the area with PC2 values around 0 to -500 and PC1 values around -1500 to 1500. The blue dots are jumping movement. They located in the area with PC2 values around -500 to -1000 and PC1 values around 0. The coordinates, Principal Components(PC1, PC2, and PC3), are linear combinations of the x, y, z coordinates of the joints. Movements that overlap in the 2D plot might be separated when considering PC3.

After performing PCA, I assigned ground truth labels to each sample and computed the centroids for each movement class in the 3D PCA space. The centroid, showed in Table 2, represents the average position of each movement type within the principal component space. Using the same

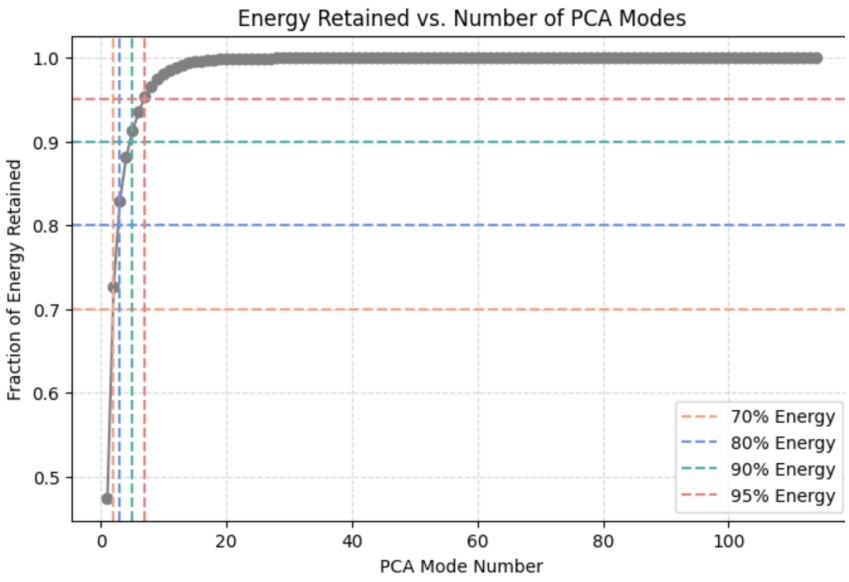

FIGURE 1. This plot shows the fraction of total energy (variance) retained in the data as a function of the number of PCA modes. The x-axis represents the number of PCA modes, while the y-axis represents the cumulative fraction of energy retained. Horizontal dashed lines indicate the energy thresholds at 70%, 80%, 90%, and 95%. The vertical dashed lines mark the minimum number of PCA modes required to achieve each energy threshold.

Centroid for class 0	-36.88211143	-253.35282541	175.91202104
Centroid for class 1	60.77197779	752.7210869	-103.41194553
Centroid for class 2	-23.88986635	-499.36826149	-72.5000755

TABLE 2. This table shows the coordinates of centers of walking(class 0), running(class 1), and jumping(class 2) data in 3D PCA space.

method to calculate the centroids with different PCA modes, I classified the training samples by computing the distance to each centroid and assigning the label of the nearest centroid.

Accuracy	k =2	k =3	k =5	k =7	k =8	k=9
Train Accuracy	88.13	75.60	75.07	87.07	87.53	87.87
Test Accuracy	98.33	92.33	91.67	94.33	93.00	94.33

TABLE 3. This table presents the accuracy rates of the classifier for different values of k (number of PCA modes) on both the training and test datasets. The training accuracy reflects the classifier's performance on the data used to build the model, while the test accuracy measures its ability to generalize to unseen data.

To evaluate the classification performance, I trained a model on the PCA-transformed data and computed both the training and test accuracy rates. The training accuracy measures how well the model fits the training data, while the test accuracy evaluates its generalization ability on unseen data. From my results shown in Table 3, k = 2 is the optimal number of PCA modes for this classification task, as it achieves the highest train and test accuracy. Higher values of k(3, 5, 7, 8, 9) tend to reduce the accuracy rate a little bit and the accuracy rates change slightly. The results indicate that a few PCA modes are enough to capture the most information of the original data

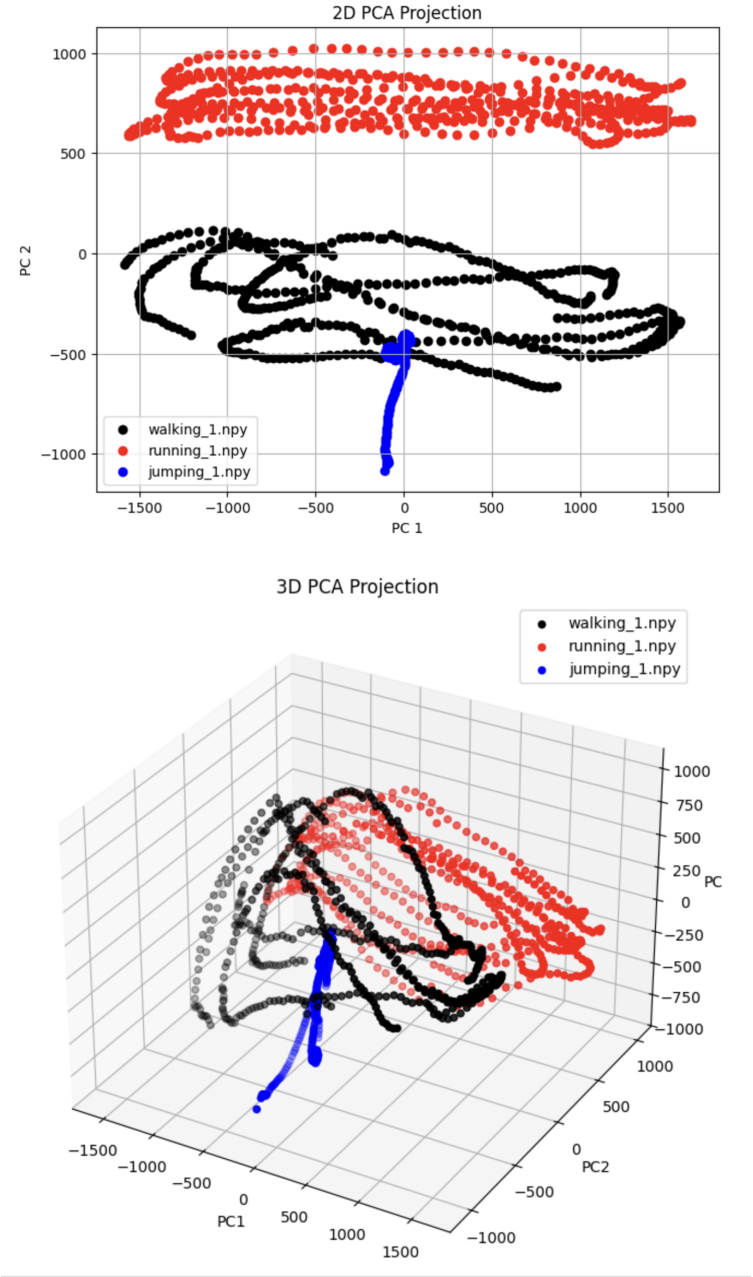

FIGURE 2. The first plot is the projection of movement data points onto two principal components: PC1(Principal Component 1) and PC2(Principal Component 2). The second plot is the projection of data points on three principal components: PC1, PC2, and PC3. There are 100 time steps for walking, running, and jumping data.

and classify data with good accuracy. More numbers of modes are not necessarily useful. Beside, the test accuracy rate are higher then the training accuracy rates. This might be the result of relatively less data in the test file.

5. Summary and Conclusions

In this assignment, I worked on the problem of captured motion recognition for a humanoid robot, OptimuS-VD, using Principal Component Analysis (PCA) and a distance-based classifier. The goal was to analyze and classify three types of movements, walking, running, and jumping, based on joint movement data recorded by the robot's sensors. This assignment demonstrated the effectiveness of PCA for dimensionality reduction and motion recognition in a humanoid robot. By reducing the data to just 2 PCA modes, I was able to achieve high classification accuracy.

Acknowledgements

The author is thankful to Dr. Frank for instructing the SVD and PCA concepts and coding with the textbook[1].

References

[1] J. N. Kutz. *Data-driven modeling & scientific computation: methods for complex systems & big data*. OUP Oxford, 2013.