RAW ACCELEROMETERY DATA ANALYSIS

Project Report

Presented to

Prof. Peter Gao

Department of Mathematics and Statistics

San Jose State University

In Partial Fulfillment

Of the Requirements of Class

SPRING 2024: MATH 250

By

Devam Sanjay Sheth

May 2024

Introduction

In recent years, the adoption of wearable accelerometer devices has gained significant traction within the realm of public health research focused on physical activity assessment. These wearable devices offer an objective and non-invasive approach to measure an individual's health levels, in contrast to the widely employed subjective methods, such as self-report questionnaires. Although subjective and objective techniques may yield comparable qualitative findings concerning factors like age and gender, the determination of adherence to activity guidelines based on accelerometer data tends to be substantially lower when compared to self-reported measures.

This project focuses on analyzing readings from accelerometer of 32 adults while they carried out different outdoor activities like walking, driving and climbing stairs. This analysis will provide us with useful insights regarding the different body part movements and help in learning the body reactions to those activities. The goal of this project is to also identify the relations between demographic data of the users with their corresponding outdoor activities.

Accelerometry data is transformed with the help of Short-term Fourier Transformer (STFT) to capture signals and frequencies with time. Variance of data is studied with the help of Principal Component Analysis (PCA). The project also focuses on application of other dimensionality reduction algorithms such as Isomap, t-SNE and Linear Discriminant Analysis (LDA). Though, the findings from them were not enough to provide full analysis, the project intends to focus more on the interpretations which can be achieved from visualizations of earlier mentioned algorithms.

Methodologies

The dataset used for this project is longitudinal and hence, a variety of algorithms can be applied in different ways. This project focuses on four different algorithms:

1. Short-term Fourier Transform (STFT): a mathematical technique used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time.

It provides a way to see how different frequencies in a signal vary at different moments, which is particularly useful for signals whose properties evolve over time, like music or speech.

STFT works by dividing a longer time signal into shorter segments of equal length. Each segment is multiplied by a window function (like a Hamming window). After applying the window, a Fourier Transform is performed on each segment. The Fourier Transform converts each piece from the time domain into the frequency domain. This shows which frequencies are present in that segment and their intensities.

- Formula of STFT is as follows:

$$X(\tau, \omega) = \int x(t)w(t-\tau)e^{\{-j\,\omega t\}dt}$$

, where $w(t-\tau)$ is a window function

The results of the Fourier Transforms are then typically represented in a spectrogram, which is a visual representation of the spectrum of frequencies as they vary with time. Each point in the spectrogram shows the intensity of a particular frequency at a specific time.

2. Principal Component Analysis (PCA): The key concept in PCA is to find a linear combination of features that maximizes the variance between classes while minimizing the within-class scatter.

It is a linear dimensionality reduction algorithm that projects data onto a new set of orthogonal axes (principal components) that capture the most variance in the data. It does not utilize class labels during transformation. PCA focuses on global structure preserving the overall data distribution in the lower-dimensional space. It is computationally efficient well-established algorithm with efficient implementations.

3. Isomap: Isomap stands for "Isometric Mapping". It is a non-linear dimensionality reduction technique. It aims to preserve the intrinsic geometric structure of high-dimensional data in a lower-dimensional space.

Isomap excels at capturing non-linear relationships in the data, which is essential when linear methods like PCA fall short. It is a manifold learning algorithm that unravels underlying low-dimensional structure (manifold) in high-dimensional data. Thus, it is a nonlinear dimensionality reduction that captures complex, non-linear relationships between data points.

4. t-SNE (t-distributed Stochastic Neighbor Embedding): t-SNE is a popular dimensionality reduction technique used in machine learning and data visualization. It is particularly effective at reducing high-dimensional data into a lower-dimensional space while preserving the local structure and relationships between data points.

t-SNE gives you a feel and intuition on how data is arranged in higher dimensions. It is often used to visualize complex datasets into two and three dimensions. t-SNE is a nonlinear technique that focuses on preserving the pairwise similarities between data points in a lower-dimensional space.

It initializes with Stochastic neighbor embedding leveraging probabilistic techniques to project high-dimensional data into a lower-dimensional space. It also focuses on local structure and preserves similarities between close data points in the high-dimensional space.

Thus, it helps in capturing non-linear relationships and local clusters well. It is computationally expensive and can be slower than PCA due to its iterative optimization process.

5. Linear Discriminant Analysis (LDA): A statistical method used for dimensionality reduction and classification tasks. It aims to find a linear combination of features that maximally separates different classes in the data.

It is also known as Fisher's Linear Discriminant. LDA helps us identify the most relevant features that contribute to class separation, improving the accuracy of classification models. It is an approach used in supervised learning to solve multi-class classification problems. LDA separates multiple classes with multiple features through data dimensionality reduction. It is generally followed by PCA, in case if data is having singular total within-class scatter matrix.

Data Description

The dataset comprises of raw labelled accelerometer readings captured from 32 healthy individuals while they were walking outdoors, climbing stairs, and driving. The

accelerometer data was simultaneously recorded at four body positions: the left wrist, left hip, left ankle, and right ankle, with a sampling rate of 100 Hz.

The dataset incorporates labels indicating the specific activity performed (walking, descending stairs, ascending stairs, driving, clapping) at each time point during the data gathering process. Additionally, basic demographic details about the participants have been included. All data has been anonymized to protect individual privacy.

This project includes raw accelerometry data files, a data files dictionary, and participant demographic information. All data are anonymized. Specifically, the project files include:

1. raw_accelerometry_data: a directory with 32 data files in CSV format. Each file corresponds to raw accelerometry data measurements of 1 study participant. File names follow the convention: "subj_id.csv". Each file contains 14 variables:
    a. activity: Type of activity (1=walking; 2=descending stairs; 3=ascending stairs; 4=driving; 77=clapping; 99=non-study activity)
    b. time_s: Time from device initiation (seconds [s])
    c. lw_x: Left wrist x-axis measurement
    d. lw_y: Left wrist y-axis measurement
    e. lw_z: Left wrist z-axis measurement
    f. lh_x: Left hip x-axis measurement
    g. lh_y: Left hip y-axis measurement
    h. lh_z: Left hip z-axis measurement
    i. la_x: Left ankle x-axis measurement
    j. la_y: Left ankle y-axis measurement
    k. la_z: Left ankle z-axis measurement
    l. ra_x: Right ankle x-axis measurement
    m. ra_y: Right ankle y-axis measurement
    n. ra_z: Right ankle z-axis measurement
2. raw_accelerometry_data_dict.csv: a CSV file containing the description of 14 variables that each file in the raw_accelerometry_data directory consists of.
3. participant_demog.csv: a CSV file with participants demographic information. The file contains 7 variables:
    a. subj_id: Participant ID (a character scalar). The value in this column can be matched with a file name (without ".csv" extension) of a file in raw_accelerometry_data directory.
    b. gender: Participant gender (a character scalar; one of: "male", "female").
    c. age: Participant age (an integer scalar).
    d. height_in: Participant height (an integer scalar; expressed in inches).
    e. weight_lbs: Participant weight (an integer scalar; expressed in pounds).
    f. race: Participant race (a character scalar; one of: "asian", "black", "caucasian").
    g. right_handed: Participant handedness (an integer scalar; 1 if right-handed, 0 otherwise).

Application and Results

Recent advancements in technology, coupled with the reduced cost of wearable devices, have significantly boosted the popularity of wearable technology in health research. Wearable physical activity (PA) monitors, particularly, hold great promise for health studies. These devices significantly differentiate among different PA into active and sedentary states, quantifying the duration of activity at various intensity levels, and accurately identifying specific types of activities such as walking, stair climbing, and driving. We explore the challenges and opportunities presented by analyzing labeled raw accelerometry data below.

i. STFT

STFT was applied to an accelerometry data of an individual, to analyze the change in frequencies from the readings with time when the candidate carries out different activities. As dataset provided with acceleration in three directions, it was time taking and compute exhaustive task to calculate STFT.

Thus, accelerations of each accelerometer were combined and only magnitude of acceleration was taken into consideration. This led to loss of information regarding the rotational movement and directions. Though this provided us with useful insights.

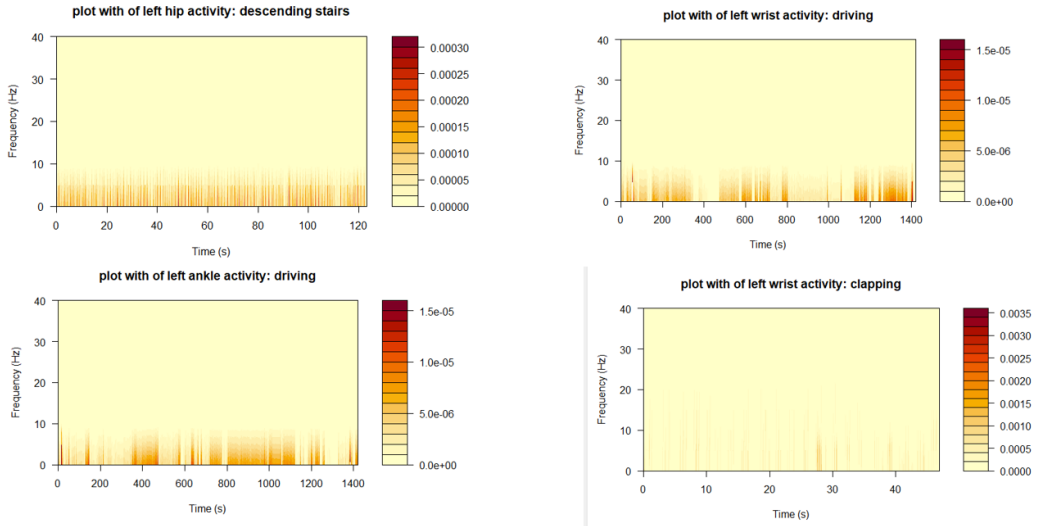

Figure 1 STFT Plots

Inferences:

1. From the above graph, contribution of different frequencies with time can be observed.
2. These visuals provide change in acceleration with time during different activities.
    a. Example: plot of left hip during descending stairs shows the periodic motion while moving down

Before moving for further analysis, demographic data of users and their corresponding accelerometer data is combined into one data frame and time column is removed. Demographic data mostly is used for categorical identification.

ii. PCA

As data was vast and longitudinal, it is a better practice to identify the variables which create the maximum variance in the data and use it to reduce dimensions as and when required. PCA being one of the most prominent tools to identify the variance, it is applied to

the combined data generated above. Initially three-dimensional data of the accelerators are considered for PCA, from which following an inference can be brought that manipulation in data can lead to loss of information (converting vector to its magnitude lost information of rotation and directions). While there were multiple PCA graphs based on combinations of activities, I considered walking and clapping for analysis as follows:

- For original data points while walking:
    - The first principal component (PC1) captures 13.8% variance
    - `la_y` and `lh_y` show significant contribution in PC1: depicting the forward movement
    - The second principal component (PC2) captures 12.6% variance
    - `lh_x` and `ra_y` have major contribution to PC2: depicting the vertical motion of the body while walking

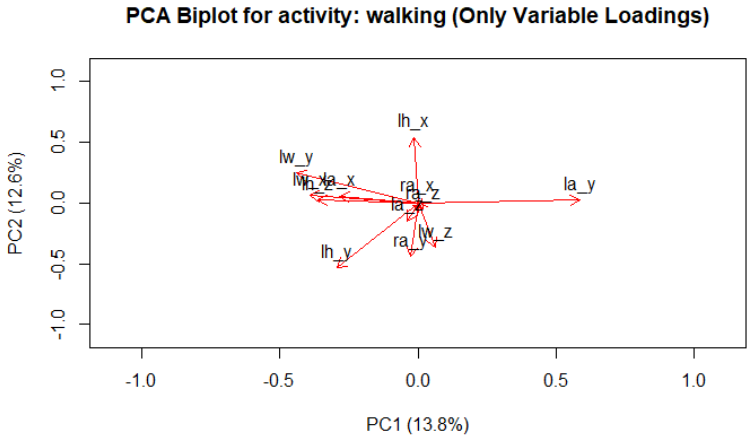

Figure 2 PCA Biplot for variable payloads

Later all the three dimensions are merged and only magnitude of accelerations are taken into consideration, which result into similar analysis. This shows that there was least variance in direction and rotation that was captured by accelerometers.

- PCA of magnitude of acceleration while walking:
    - The first principal component (PC1) captures 28.2% variance
    - Magnitude of all the accelerometers have negative contribution to PC1, but are highly correlated
    - The second principal component (PC2) captures 24.7% variance
    - Except `ra_mag` (magnitude of acceleration of right ankle), other components had least contribution

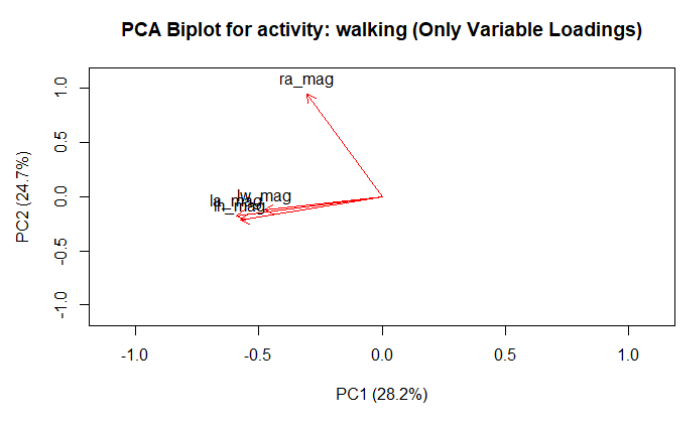

Figure 3 PCA over data containing the magnitude of the accelerations

iii. Isomap

Though PCA was able to provide my insights about variances, it was unable to provide the distinct clusters on the data. Thus, I headed towards using Isomap. Isomap being a manifold learning algorithm, it tries to open up the shape into lower dimensions following geodesic distances.

When isomap is applied over full data, it just forms one cluster without providing any distinction among activities.

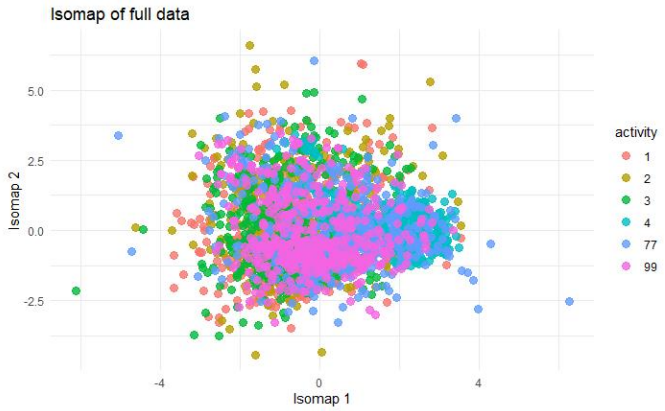

Figure 4 Isomap of full data

So, Isomap is then applied over different activities and different accelerometers separately. Following which some interesting patterns can be seen:

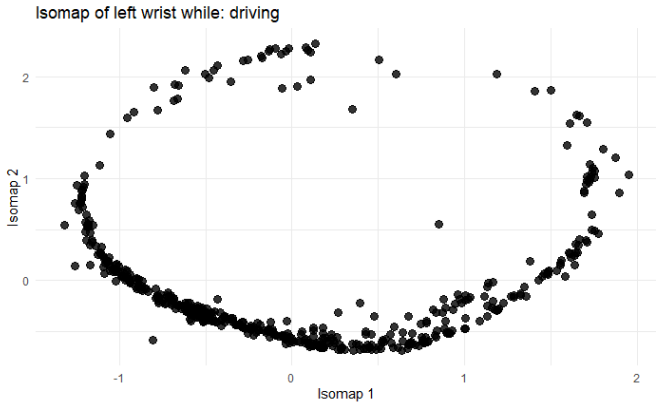

Figure 5 Isomap of left wrist while driving

Considering the above graph, it can be seen that isomap tried to open up the datapoints and it clearly shows the hand movement of left wrist while driving.

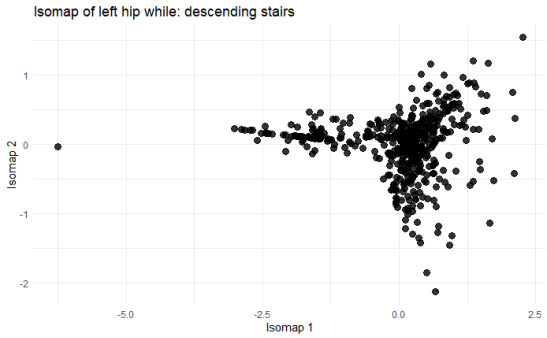

Figure 6 Isomap of left hip while descending

The above isomap graph shows the movement of left hip while descending the stairs. If seen carefully, variation along isomap1 shows the sudden jerk when left leg is put down and variation along isomap 2, shows the swing movement of hip while descending.

iv. t-SNE

Till now, there has been no clustering which could provide with a good insight and distinguish various activities. t-SNE, being a dimensional reduction algorithm which casts t-distribution over data and tries to enhance distinction, is applied over the data. tSNE is applied over three dimensional columns of each accelerator.

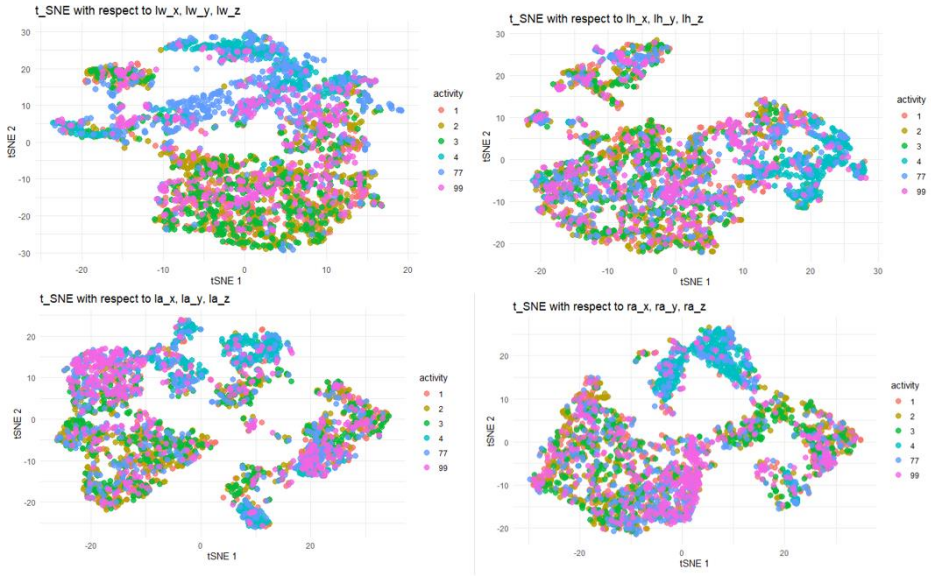

Figure 7 tSNE over whole data

On applying tSNE over data, though there was no distinction, but it can be observed that one activity had dominance over the other in clusters. Focusing on tSNE graph in bottom right, it can be seen that activity 99 (non-studied activity) has major dominance with right ankle accelerator in one cluster, whereas activity 4 (driving) has a dominance in top cluster.

Similarly, for the tSNE graph in top-right corner corresponding to left ankle accelerator, it can be seen there is a major dominance of activity 3 (ascending stairs) in lower cluster, whereas activity 4 and 77 (clapping) has major dominance in upper cluster.

v. LDA

Even after using tSNE, there were no significant clusters which could be found. LDA is a clustering algorithm. As it aims to find a linear combination of features that maximally separates different classes in the data, this can help in identifying the clusters.

Following the same technique, LDA is applied on data considering an accelerator data at a time and calculating LDA over its three-dimensional data.

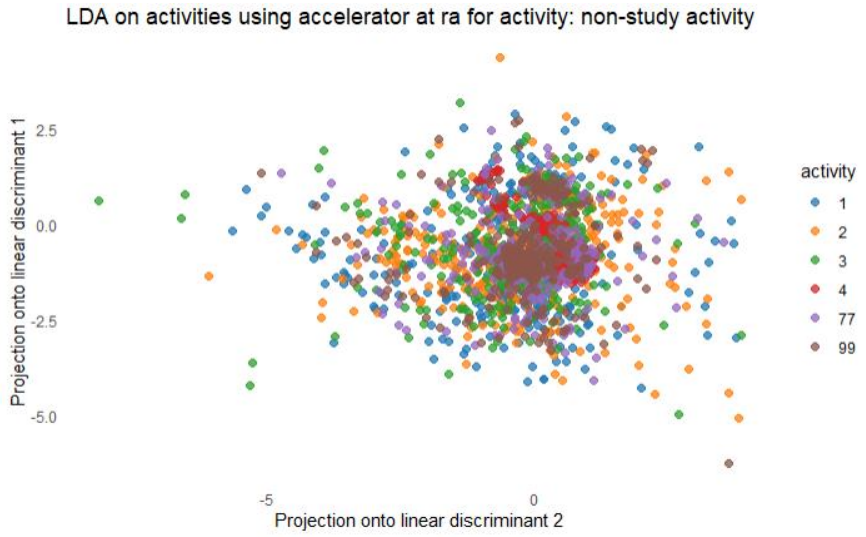

Figure 8 LDA over full data

On observing the visualization above, no distinct clusters could be found.

As part of optimization, PCA was applied before LDA. Initially the data was classified based on gender, but it was found that there was near to no difference in the graphs of both the genders. Thus, it can be concluded that data for both male and female was highly correlated.

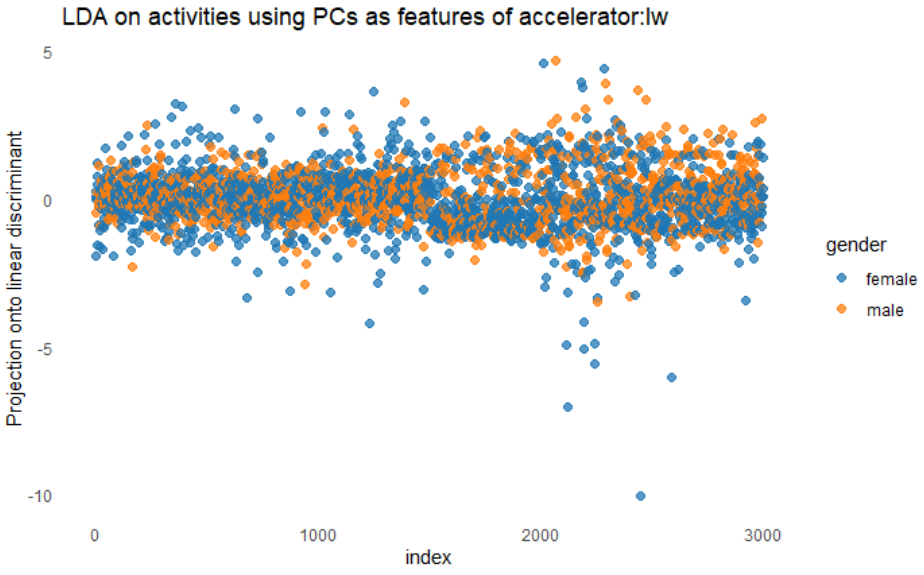

Figure 9 LDA on PCA for gender

Apart from which, same method was applied for activities, which gave a good classification, but it was classified based on index (that is position of data points in data frame).

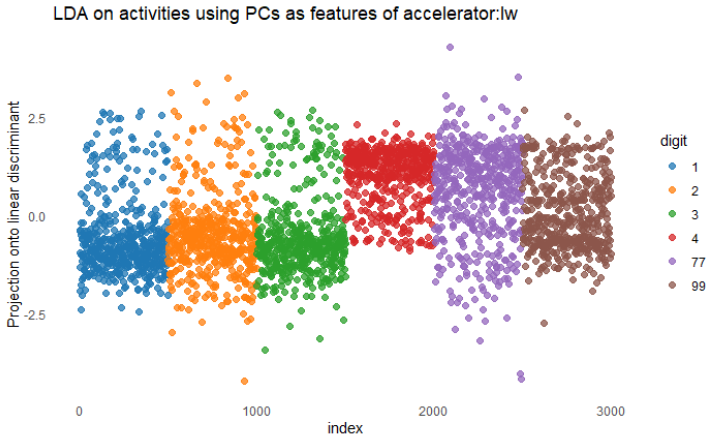

Figure 10 LDA on PCA with indices on x-axis

1. Some clustering taking place between activities 1, 2, and 3 against the cluster of 4 and 77, along LD1 for accelerator is observed.
2. Activities 1, 2, and 3 seem to be on negative projection of LD1 and highly correlated to each other, whereas activities 4, and 77 are more on positive portion of LD1.

Thus, to get some conclusive clusters, time_s variable was used, which showed when each datapoint was captured, as follows

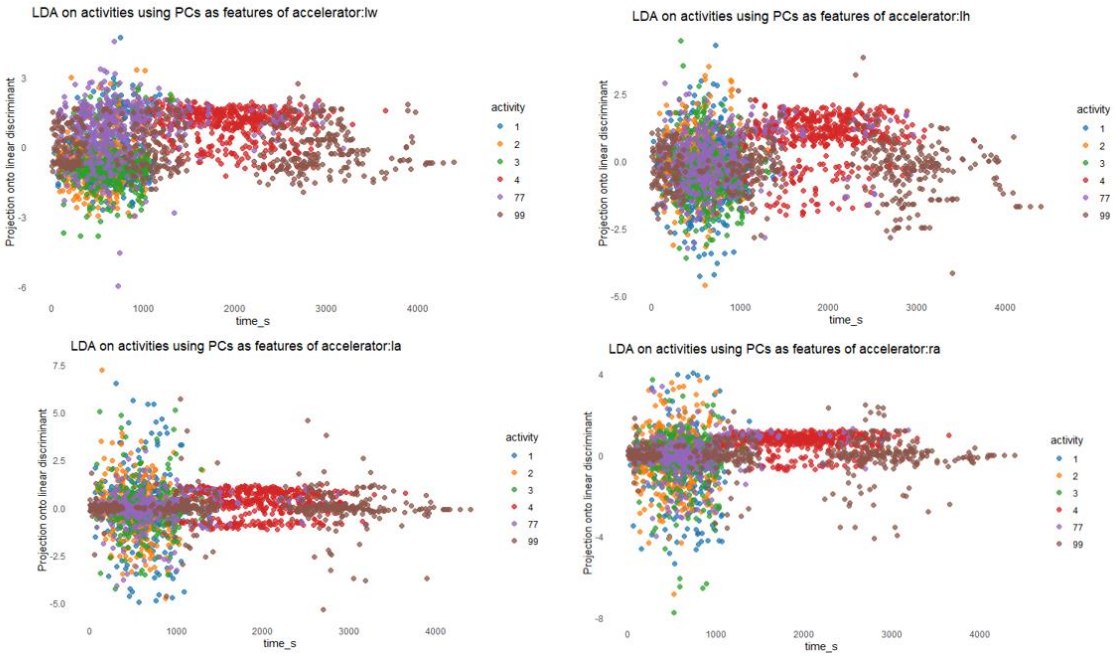

Figure 11 LDA on PCA with time on x-axis

Following points were concluded:

1. Driving activity had least variance along the time and cluttered around 0 for all the accelerators

2. Time when non-study activities were carried out also had least variance and had a gap in between as two clusters are formed with time
3. Rest all activities have high variance and were carried out in short duration.

Even projections of points on two prominent LDAs were also plotted as follows:

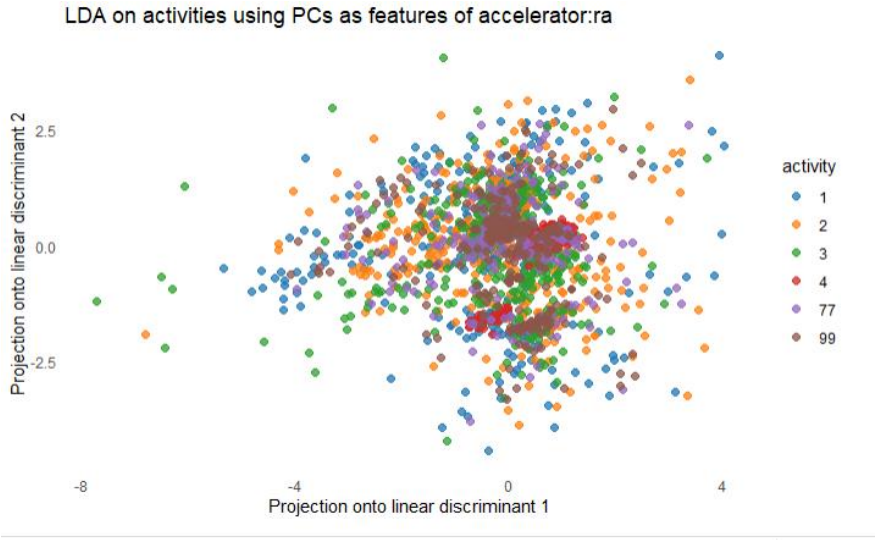

Figure 12 LDA on PCA with LD1 and LD2 as axes

But Nothing could be inferred from it.

Conclusion

- STFT can provide a good amount insight about frequencies along with time, providing a good insight about the movements
- PCA provided a with the information useful to identify relations between variables and their corresponding variances
    - Though, it also led to some over-interpretation
- Isomap gave out some patterns, but was mostly useful for category of data (driving)
- t-SNE provided with a single cluster, with no significant impactful results
- LDA provides with partial difference between activities, but not significant
    - LDA along time provides a some inference about variation of acceleration in different activities

Future Work

- Apply other dimensionality reduction and clustering algorithms like diffusion mapping
- Enhance data pre-processing
- Identify better cluster inferences

Reference papers

1. Fadel, W. F., Urbanek, J. K., Albertson, S. R., Li, X., Chomistek, A. K., & Harezlak, J. (2019). Differentiating Between Walking and Stair Climbing Using Raw Accelerometry Data. Statistics in Biosciences, 11(2), 334–354. doi: https://doi.org/10.1007/s12561-019-09241-7

2. Straczkiewicz, M., Urbanek, J., Fadel, W., Crainiceanu, C., & Harezlak, J. (2017). Automatic Car Driving Detection Using Raw Accelerometry Data. Innovation in Aging, 1(suppl_1), 1239–1239. doi: https://doi.org/10.1093/geroni/igx004.4499

3. Karas, M., Bai, J., Strączkiewicz, M., Harezlak, J., Glynn, N. W., Harris, T., … Urbanek, J. K. (2019). Accelerometry Data in Health Research: Challenges and Opportunities. Review and Examples. Statistics in Biosciences, 11, 210–237. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6874221/