# Assignment 2

Martijn Stoorvogel

## Part 1: preliminary data analysis and pre-treatment

In this part of the assignment, you will reflect on what is contained in your data and how you can pretreat your data.

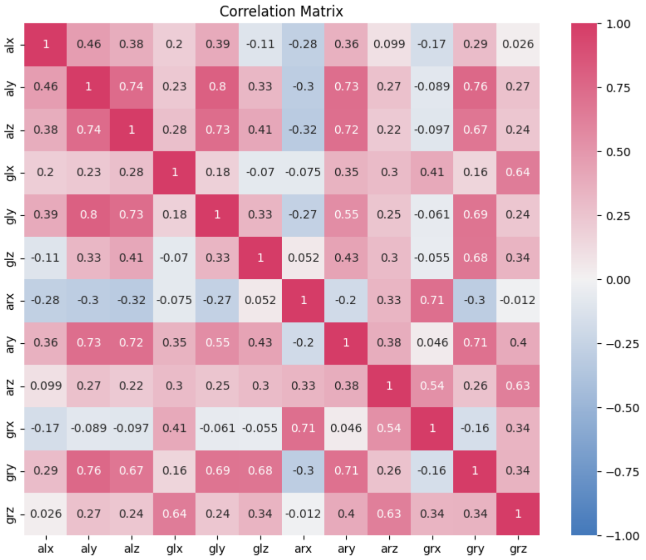
### 1.1. Preliminary analysis

**Q1.1.** Are there any highly correlated variables (*i.e.*, correlation larger than 0.90)? [2 points]

- □ Yes
- □ No

Visualize this answer with a *single* figure and report it below.

[figure here]



| | alx | aly | alz | glx | gly | glz | arx | ary | arz | grx | gry | grz |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| alx | 1 | 0.46 | 0.38 | 0.2 | 0.39 | -0.11 | -0.28 | 0.36 | 0.099 | -0.17 | 0.29 | 0.026 |
| aly | 0.46 | 1 | 0.74 | 0.23 | 0.8 | 0.33 | -0.3 | 0.73 | 0.27 | -0.089 | 0.76 | 0.27 |
| alz | 0.38 | 0.74 | 1 | 0.28 | 0.73 | 0.41 | -0.32 | 0.72 | 0.22 | -0.097 | 0.67 | 0.24 |
| glx | 0.2 | 0.23 | 0.28 | 1 | 0.18 | -0.07 | -0.075 | 0.35 | 0.3 | 0.41 | 0.16 | 0.64 |
| gly | 0.39 | 0.8 | 0.73 | 0.18 | 1 | 0.33 | -0.27 | 0.55 | 0.25 | -0.061 | 0.69 | 0.24 |
| glz | -0.11 | 0.33 | 0.41 | -0.07 | 0.33 | 1 | 0.052 | 0.43 | 0.3 | -0.055 | 0.68 | 0.34 |
| arx | -0.28 | -0.3 | -0.32 | -0.075 | -0.27 | 0.052 | 1 | -0.2 | 0.33 | 0.71 | -0.3 | -0.012 |
| ary | 0.36 | 0.73 | 0.72 | 0.35 | 0.55 | 0.43 | -0.2 | 1 | 0.38 | 0.046 | 0.71 | 0.4 |
| arz | 0.099 | 0.27 | 0.22 | 0.3 | 0.25 | 0.3 | 0.33 | 0.38 | 1 | 0.54 | 0.26 | 0.63 |
| grx | -0.17 | -0.089 | -0.097 | 0.41 | -0.061 | -0.055 | 0.71 | 0.046 | 0.54 | 1 | -0.16 | 0.34 |
| gry | 0.29 | 0.76 | 0.67 | 0.16 | 0.69 | 0.68 | -0.3 | 0.71 | 0.26 | -0.16 | 1 | 0.34 |
| grz | 0.026 | 0.27 | 0.24 | 0.64 | 0.24 | 0.34 | -0.012 | 0.4 | 0.63 | 0.34 | 0.34 | 1 |

*Figure 1: Correlation matrix of the sensor data*

**Q1.2.** Are there any missing data? [3 points]

- □ Yes
- □ No

**Q1.3.** How many duplicated rows are present in your dataset, if any? Specify the number below: [5 points]

0 duplicated rows are present in my dataset. I have checked for any duplicates as can be seen in the code.

________

*If you have replied yes to Q1.1 remove one of the highly correlated variables from your dataset. If you have found missing data or duplicates, proceed to remove the corresponding rows. If* *you have not found correlated variables, missing data, or duplicates, you can proceed to the next parts of the questionnaire and use your pretreated data for the assignment.*

## 1.2. Data scaling

When choosing whether to scale the data and what scaling procedure to use, one should (a) analyze how data distribute, and (b) reflect on the meaning of each feature.

**Q1.4.** What type(s) of scaling procedures among the ones reported below would be in principle correct for the dataset under analysis and why? Select the answer(s) you consider correct, and explain why you made your choices. [multiple choices possible! 10 points]

| **Type of scaling** (check if correct) | **Reason for your answer** |
|---|---|
| *No scaling* (data used as they are) | PCA is sensitive to the relative variances of the variables, and variables with larger scales can dominate the analysis. Scaling the variables to a similar range ensures that PCA gives appropriate weight to all variables, regardless of their original scales. |
| *Mean centering* (mean equal to 0) | Mean centering alone will only adjust the means of the variables to zero but will not address the issue of differing variances. |
| *Unitary variance* (variance equal to 1) | Unity variance will adjust the variance of all variables to one so they will be comparable on their variance. However, the mean values are not adjusted. This way values with higher means will influence the principle components more. |
| *MinMax scaling* (minimum and maximum equal to 0 and 1, respectively) | With MinMax scaling will bring all variables to a common scale while preserving their relative differences. This way larger values no longer dominate the principle components. Outliers are also accounted for by preserving the values between 0 and 1. |
| *Standard scaling* (mean equal to 0 and variance equal to 1) | With standard scaling it is made sure that biases caused by high mean values and high variances are removed from the data. It helps to compress extreme values closer to the range of the majority of the data. |

You can report *one* figure below if this helps you explain your reasoning.

[figure here]

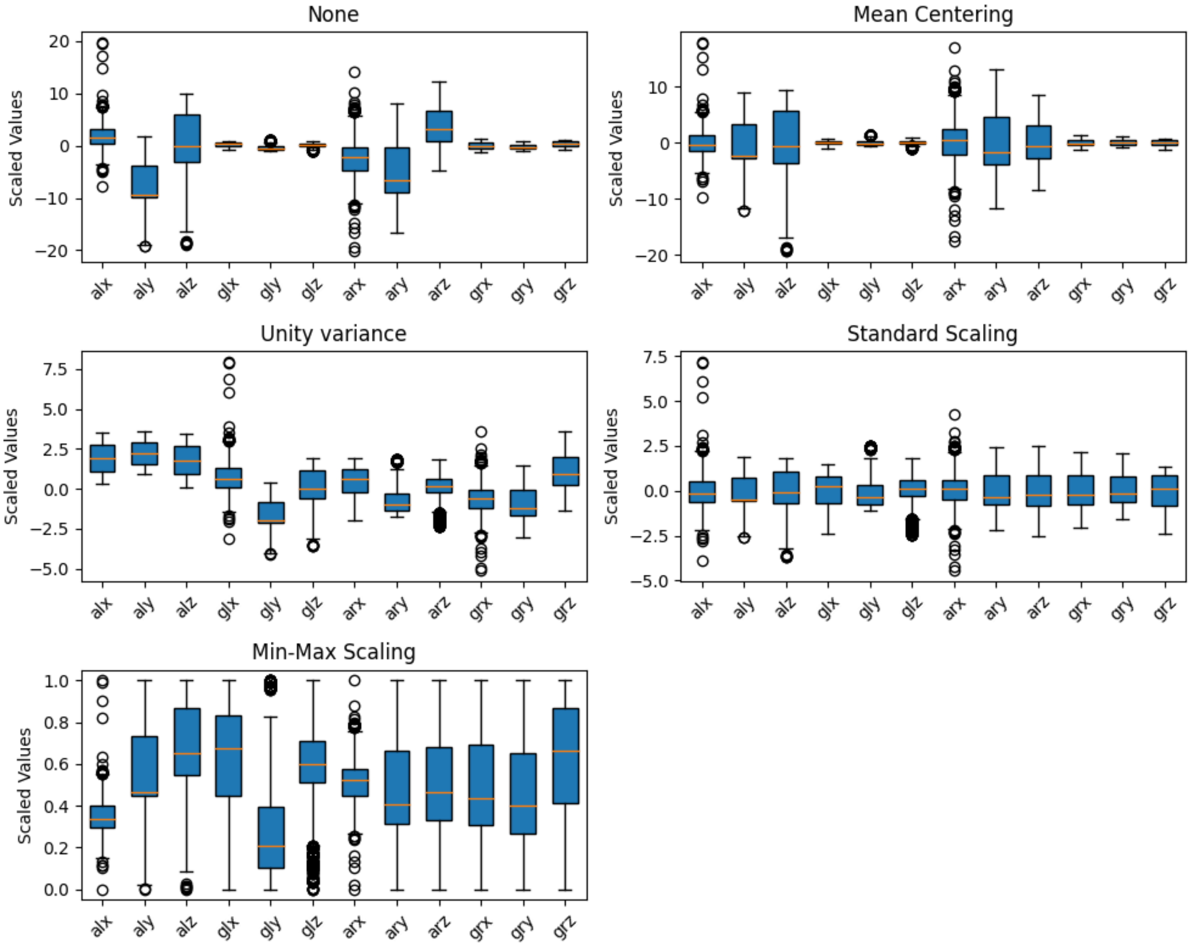


*Figure 2: Visualization of various scaling methods on the sensor data.*

## Part 2: Principal Component Analysis

In this part of the assignment, you will start performing the steps of a classical PCA analysis.

### 2.1. Selection of the number of principal components

For this exercise, take your pretreated dataset as in Part 1. To allow for comparability of the results across all of you, you will use MinMax scaling.

**Q2.1** How many variables are necessary to capture at least 90% of your dataset variance (given the steps explained above)? Insert your answer below: [5 points]

6 components are needed to at least cover 90% of the dataset variance as can be seen in the plot below.

______

Visualize your answers with a figure and report it below.

[figure here]

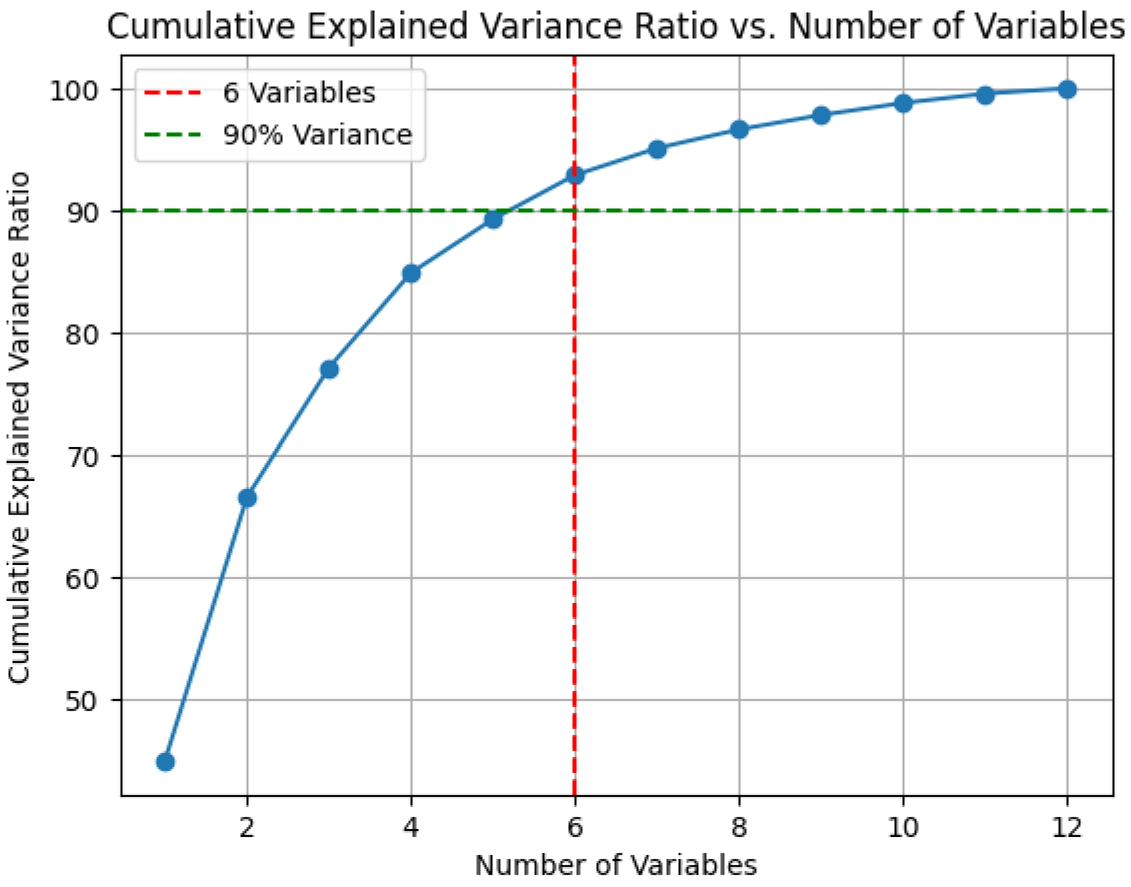


*Figure 3: The cumulative explained variance of the principle components*

*For the rest of the exercise, we will be using the number of principal components you have reported in your answer. Be mindful: if your components are too many to solve this exercise swiftly, something might have gone wrong in the pipeline ;)*

### 2.2. Analysis of the variables

**Q2.2** What variable has the largest effect on PC1? [5 points]
gry

Motivate how you reached this conclusion below (maximum 100 words).
The loadings represent the correlation between the principle component and the variable. Therefore the variable with the largest loading in PC1 is the variable that has the most effect. In this case it was gry with a loading of 0.466.

**Q2.3** What variable has the largest effect on PC2? [5 points]
grx

Motivate how you reached this conclusion below (maximum 100 words).
Again with the same reasoning that the variable with largest loading in that principle component has the largest effect. In this case that is grx with 0.521

**Q2.4** Is there a variable that is not relevant to compute PC2? [5 points]

- □ [✓ Yes]
- □ No

Motivate your answer (max 50 words)
The contribution of variable alx is almost 0 for PC2, which indicates orthogonality for that variable with the current principle component. Thus it has almost no influence on PC2

**Q2.5** Do variables coming from the same sensor contribute similarly to PC1 and PC2? [10 points]

- □ Yes, in all cases
- □ Yes, in some cases
- □ [✓ No, with few exceptions]
- □ No, never

It only is the case for alx, however the contribution in both principle components is very low/negligible. For the ones that are contributing in the one, they contribute less in the other usually.

Visualize your answers with a *single* figure and report it below.

[figure here]

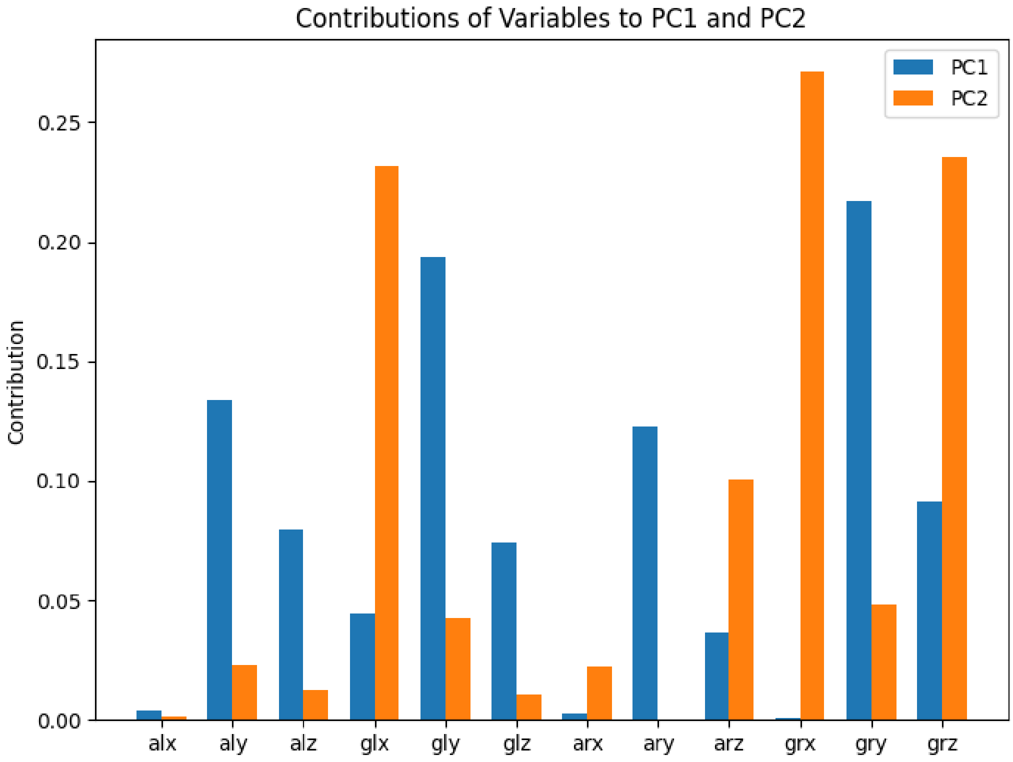


*Figure 4: Contribution of the principle components for both PC1 and PC2.*

Analysis of the samples

Have a look at the score plots obtained in the first two principal components, and answer to the questions below.

**Q.2.6** What does each point in your PCA score plot represent? [5 points]

- □ The measurements of a single sensor across subjects, activities, and timeframes
- □ A single timeframe across subjects
- □ A snapshot of a subject in a timeframe, while performing a certain activity
- □ A subject across timeframes, while performing a certain activity
- □ A snapshot of a subject in a timeframe, while performing multiple activities
- □ Other, specify: ___

**Q.2.7.** Do your samples form somewhat distinct groups in the space of PC1 and PC2? [5 points]

- □ Yes
- □ No

## Part 3: Interpretation

Here, you will use the insights you gathered so far, to interpret the data.

**Q.3.1.** If you have answered *yes* to the previous question (Q.2.7), do these groups correspond to some information you have available, and if so, what information? [10 points] (Hint: consider colouring the points in your score plot using different types of information)

- □ There are no evident groups.
- □ They do not correspond to any information available.
- □ The sample group according to the row ordering in the original dataset.
- □ The groups correspond to the difference between study subjects.
- □ The groups correspond to certain activity types.
- □ The groups capture the passing of measuring time.

Motivate your answers to **Q.2.9** and **Q.3.1** with a *single* figure and report it below.
[figure here]

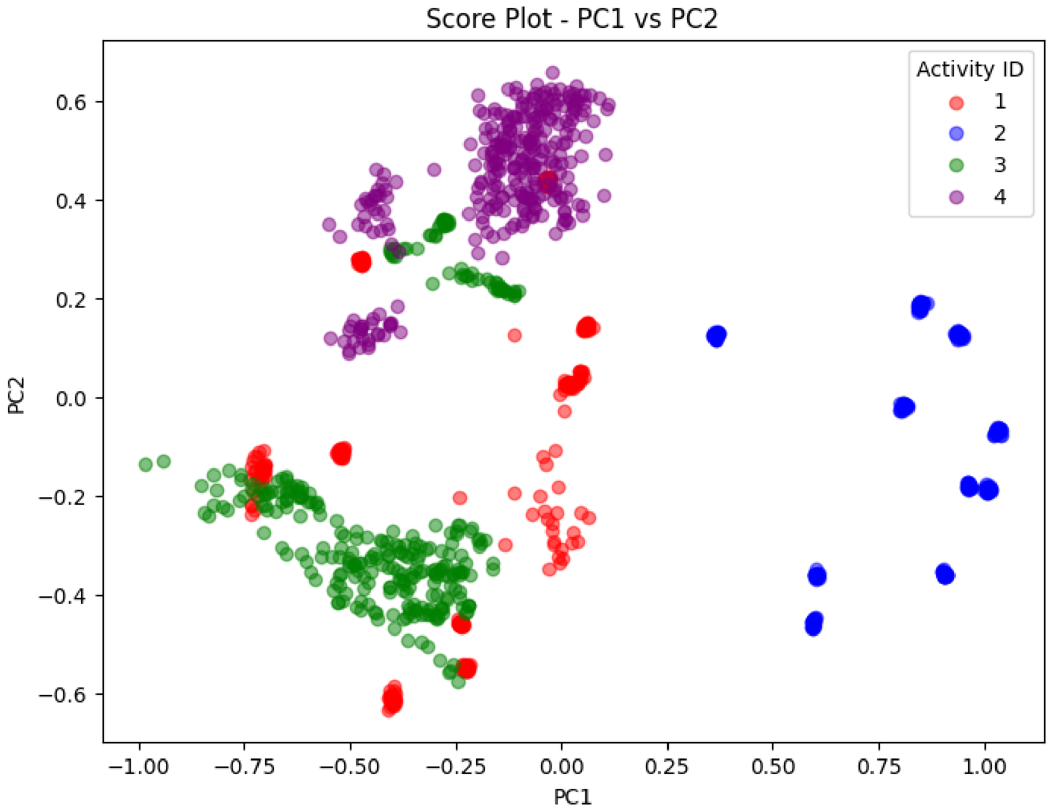


*Figure 5: Score plot of PC1 vs PC2*

**Q.3.2.** Are there features that clearly separate class 1 from the other ones? [10 points]

- □ Yes
- □ No
- □ Impossible to say using only these two components

Write some text and use a figure to explain your reasoning. If you answered 'yes', provide a list of the features you identified to support your reasoning (max 100 words).
From the plot it can be seen that using the first two principle components the group belonging to class 1 is somewhere in the middle of the plot. This explains that the variance of that group is not described by the first two principle components yet. However it could be that another principle component does capture this. To give a definitive answer as to whether the features can separate class 1 from the others the other principle components should also be analyzed.

[figure here]

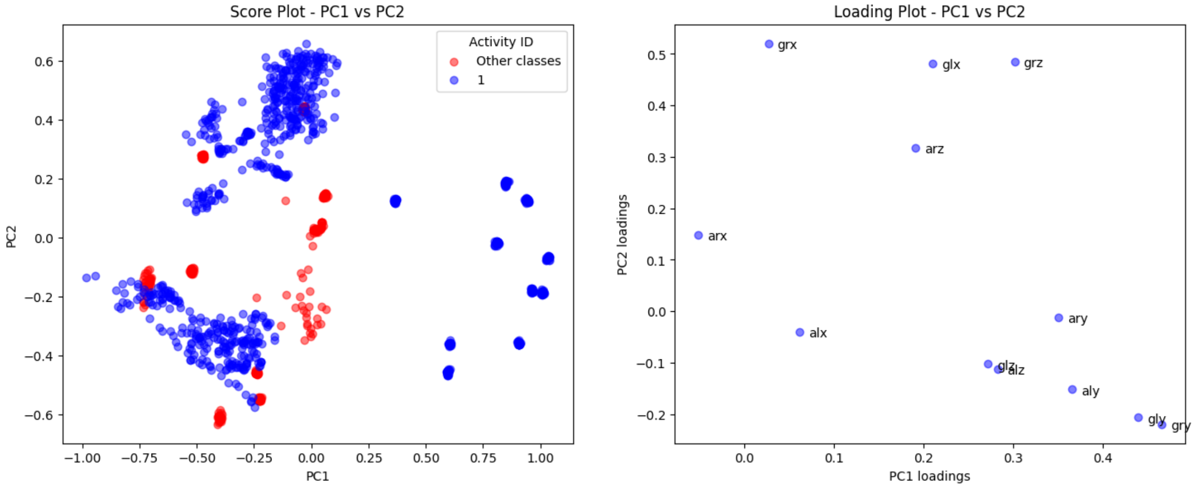


*Figure 6: Score plot and corresponding loading plot of PC1 vs PC2 highlighting class 1.*

**Q.3.3.** Are there features that clearly separate class 2 from the other ones? [10 points]

□ Yes

□ No

□ Impossible to say using only these two components

Write some text and use a figure to explain your reasoning. If you answered ‘yes’, provide a list of the features you identified to support your reasoning (max 100 words).

The variance of class 2 is captured by the first two principle components as we can see that the larger PC1 is the more likely it belongs to class 2. The features that describe this are all features found on the bottom right in the loadings plot [glz, alz, ary, aly, glx, gry].

---

[figure here]

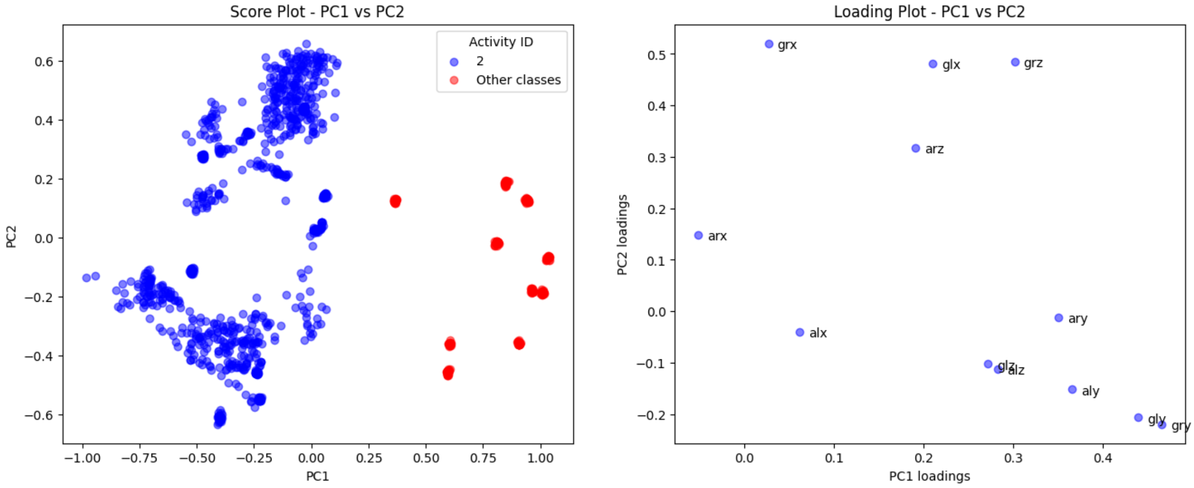


*Figure 7: Score plot and corresponding loading plot of PC1 vs PC2 highlighting class 2.*

**Q.3.4.** If one were to develop an approach to distinguish only between Walking (activity no. 2) and Jogging (activity no. 4) only, and had to choose between measuring only acceleration or gyroscopic information, what would they choose? [10 points]

□ gyroscopic data
□ acceleration data

Write some text and use a figure (optional) to explain your reasoning (max 100 words).
To separate the activity jogging and walking from each other we need to find a clear boundary. The groups can be separated using PC2, since the walking group has higher values for PC2 generally than jogging. Furthermore, also PC1 captures the difference as jogging is more to the right than walking in this direction. To capture both differences in PC1 and PC2 the gyroscopic data should be used. The data from the gyroscopic sensors is more spread in the PC1 and PC2 direction than the acceleration sensor data. Basically the gyroscopic sensor data forms the groups jogging and walking itself already.

[figure here]

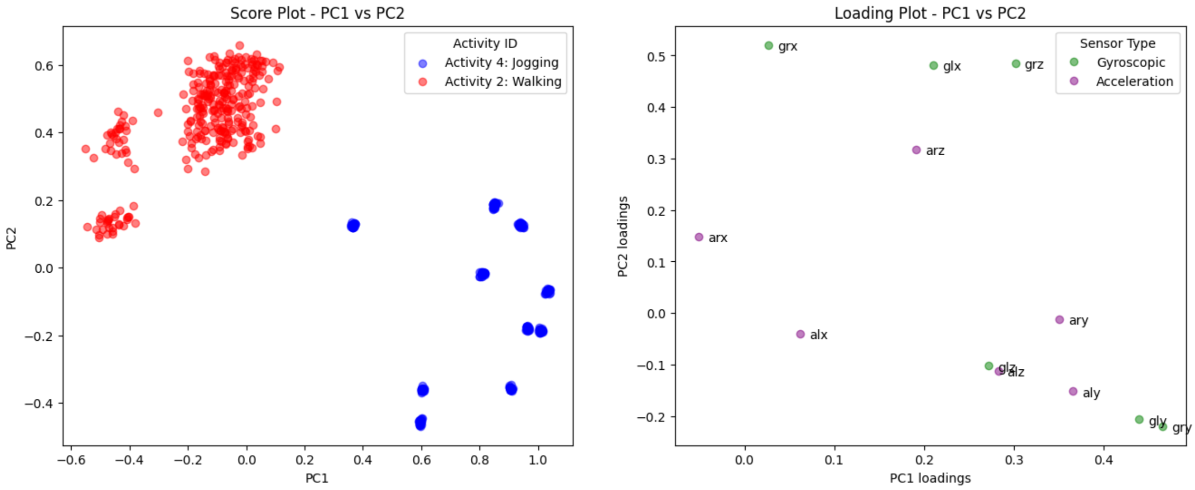


*Figure 8: Score plot and corresponding loading plot highlighting both class 2 and 4.*

If you had to attempt an interpretation of what you just noted with your data analysis, what would it be? (max 300 words) [bonus question, it adds up to 10 extra points if your answer is interesting and well-thought of]

[figure here]

Selecting a subset of principal components with the largest eigenvalues can lead to a reduction in dimensionality of the data while retaining the most valuable information. In this case, the amount of data needed to seperate the 4 activity groups could be reduced from 12 variables to 6 already while preserving the ability to seperate the groups. This simplifies the data analysis and visualization hilst preserving the essential patterns to seperate the groups from each other.