Assignment 2

Bob Verbeek (1722510)

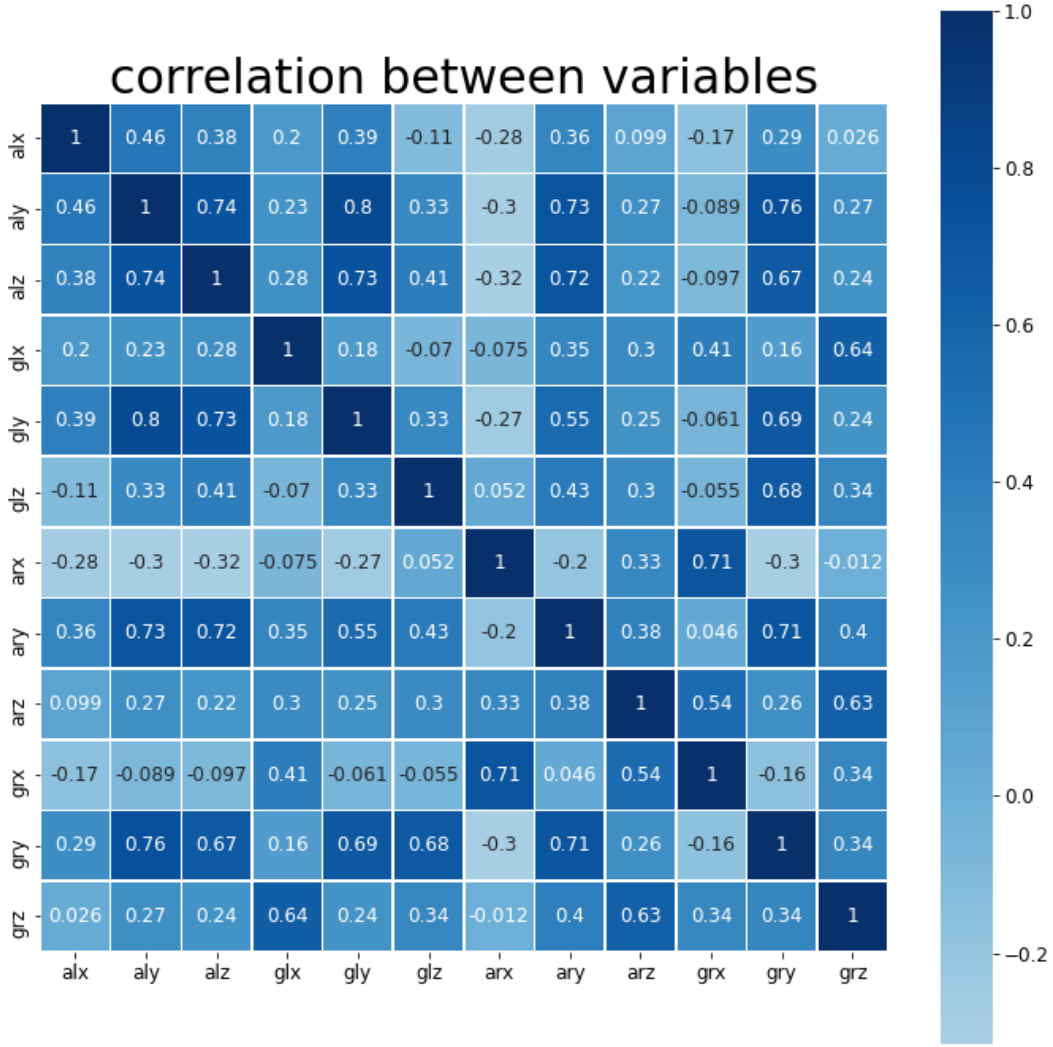
Part 1: preliminary data analysis and pre-treatment

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]

- □ No



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

- □ No the function: original_df.dropna().empty coulnd't find any empty cells.

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

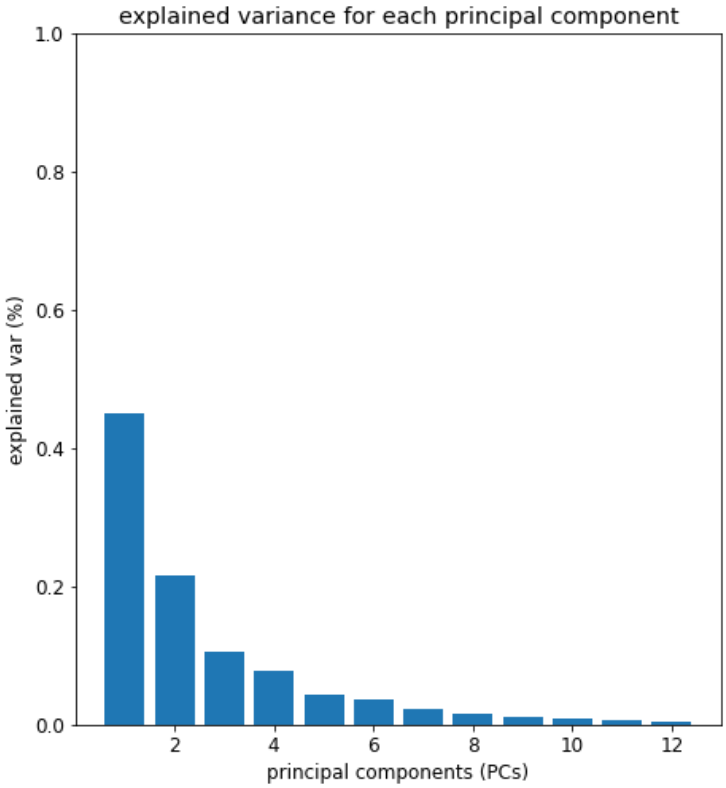
None the function: original_df[original_df.duplicated()==True] didn't find anything.

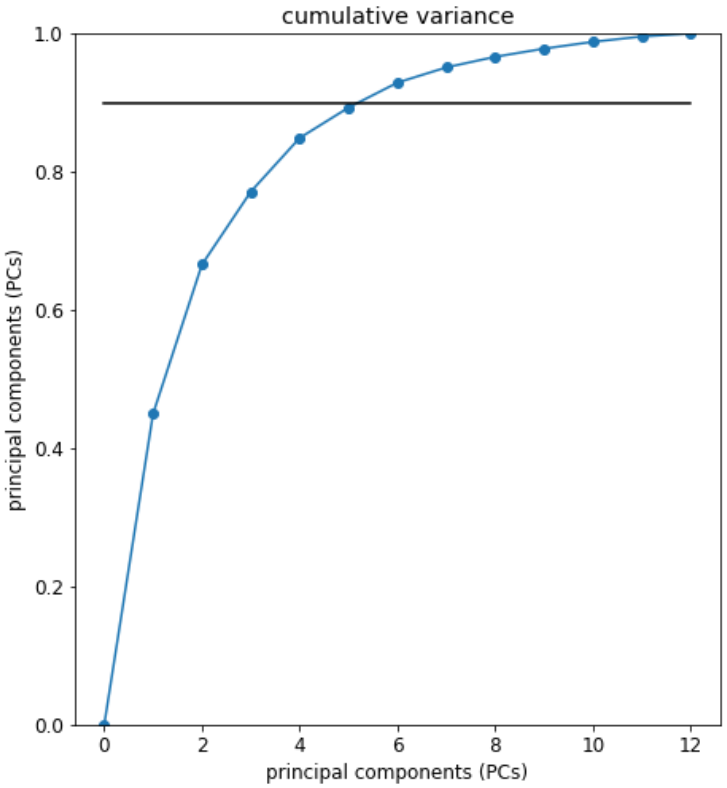
1.2. Data scaling

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) Incorrect for this dataset	No scaling is possible if your data is already on the same scale, but if not certain variables will have more influence due to higher values. This leads to bias and not to good data analysis.
Mean centering (mean equal to 0) Incorrect for this dataset	Centering the mean at 0 may be useful in some datasets, but it doesn't really scale the data so these variables will still have a larger impact and this is thus not a good type to use for this dataset.
Unitary variance (variance equal to 1) Possible and correct for this dataset	This is a useful method that is usable on the dataset since setting the variance to 1 will scale the data.
MinMax scaling (minimum and maximum equal to 0 and 1, respectively) Possible and correct for this dataset	MinMax scaling is very useful as it scales the data between 0 and 1, it also normalises the data. This makes is overall very useful to make sure the dataset is properly analysed.
Standard scaling (mean equal to 0 and variance equal to 1) Possible and correct for this dataset	This is very similar to unitary variance, but it also makes the mean equal to 0. Due to the making the variance 1 the data is scaled, but because the combination with setting the mean to 0 the distribution changes and thus the dataset. The data is still properly scaled and thus it can be used to do proper data analysis.





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

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 PCs are needed to capture at least 90% of the dataset, since 5 is just below the 90%.

Visualize your answers with a figure and report it below.

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]

Variable gry

Motivate how you reached this conclusion below (maximum 100 words).

The loading determines the contribution of every variable on a PCs, gry contains the max value for the column of PC1 and thus contributes the most towards the value of PC1. To obtain this the loadings are calculated with pca.components_.T, which transposes the matrix of loadings. The labels and indexes are then added and the new dataframe is ready.

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

Variable grx

Motivate how you reached this conclusion below (maximum 100 words).

The same methods were used as in question 2.2 with one minor change, the column was changed to PC2 instead of PC1.

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

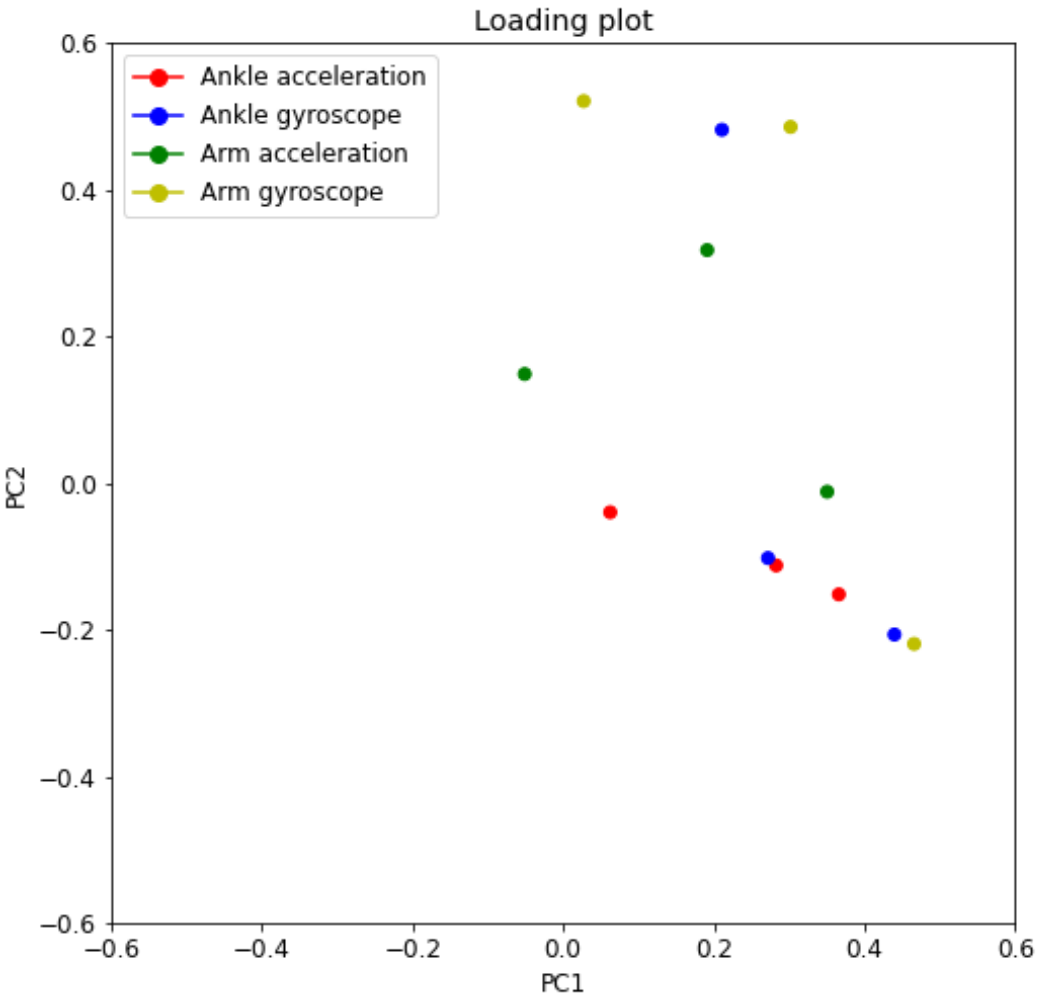
- □ Yes

Motivate your answer (max 50 words)
Multiple variables have loading values below 0.1 for pc2, these have such low values they can be seen as insignificant. These variables are alx and ary.

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

- □ No, with few exceptions

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



2.3. 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]

- □ A snapshot of a subject in a timeframe, while performing a certain activity

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

- □ Yes

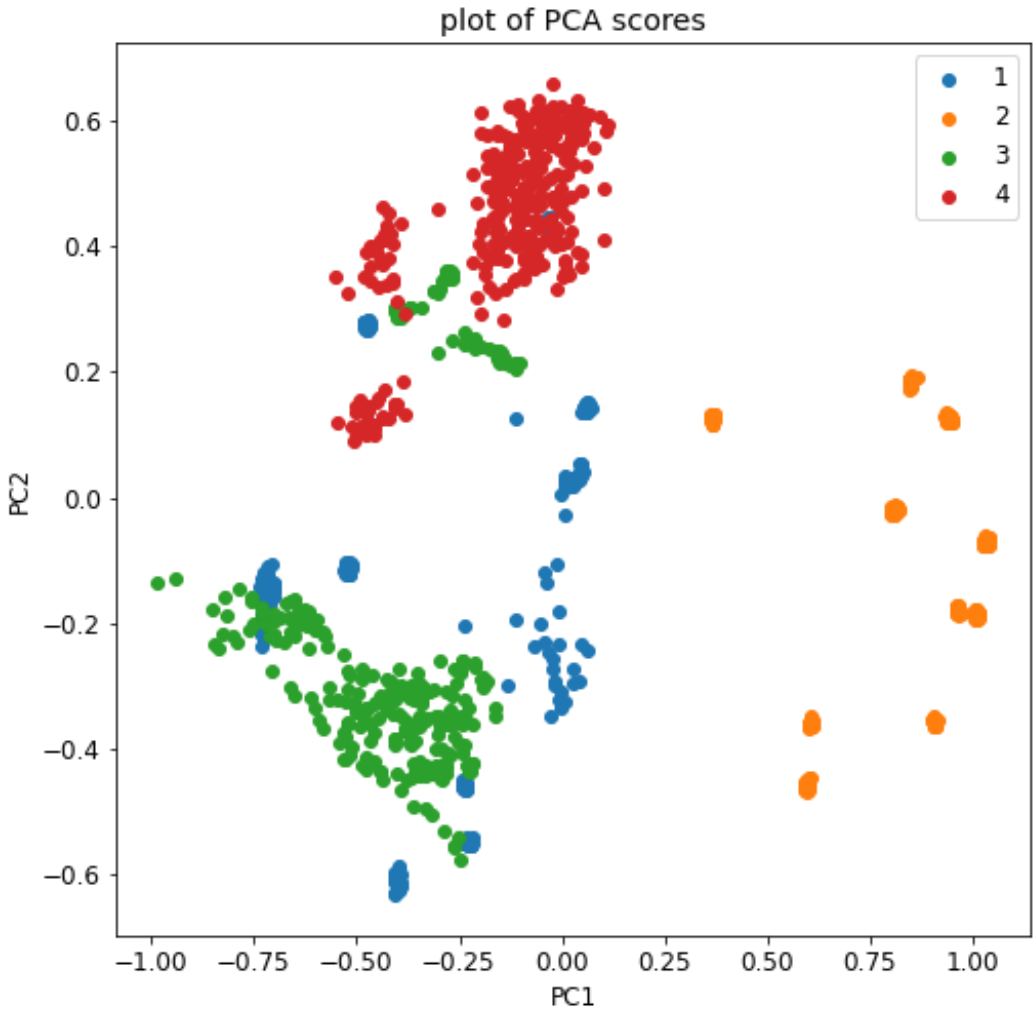
Part 3: Interpretation

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)

- □ The groups correspond to certain activity types.

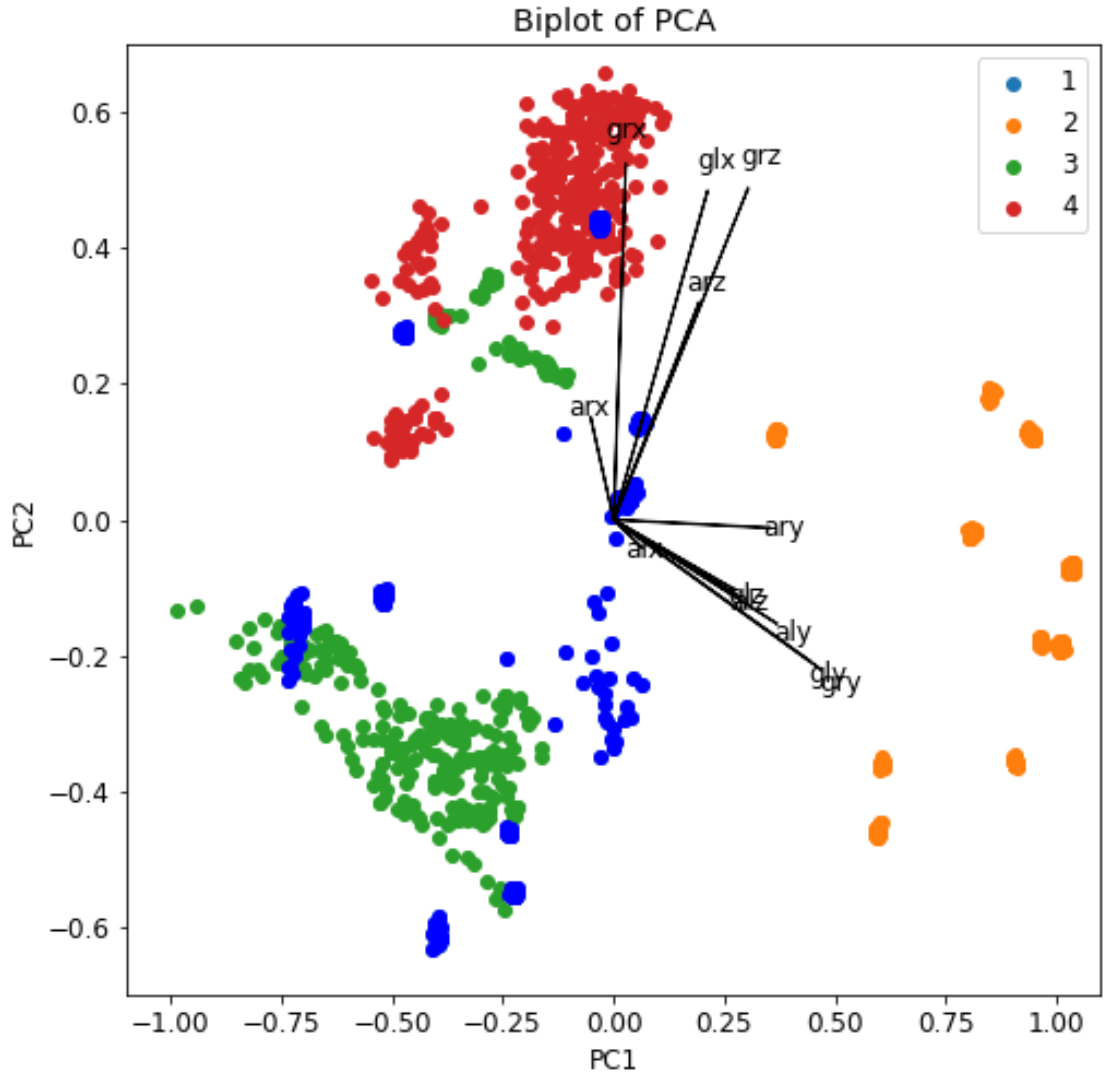
Motivate your answers to **Q.2.9** and **Q.3.1** with a *single* figure and report it below.



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

- □ 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).
It's hard if not impossible to say due to class one being loosely grouped, therefore having a lot of outliers and not being oriented in a clear direction. This makes it difficult to determine which if any features clearly separate class 1 from the others.

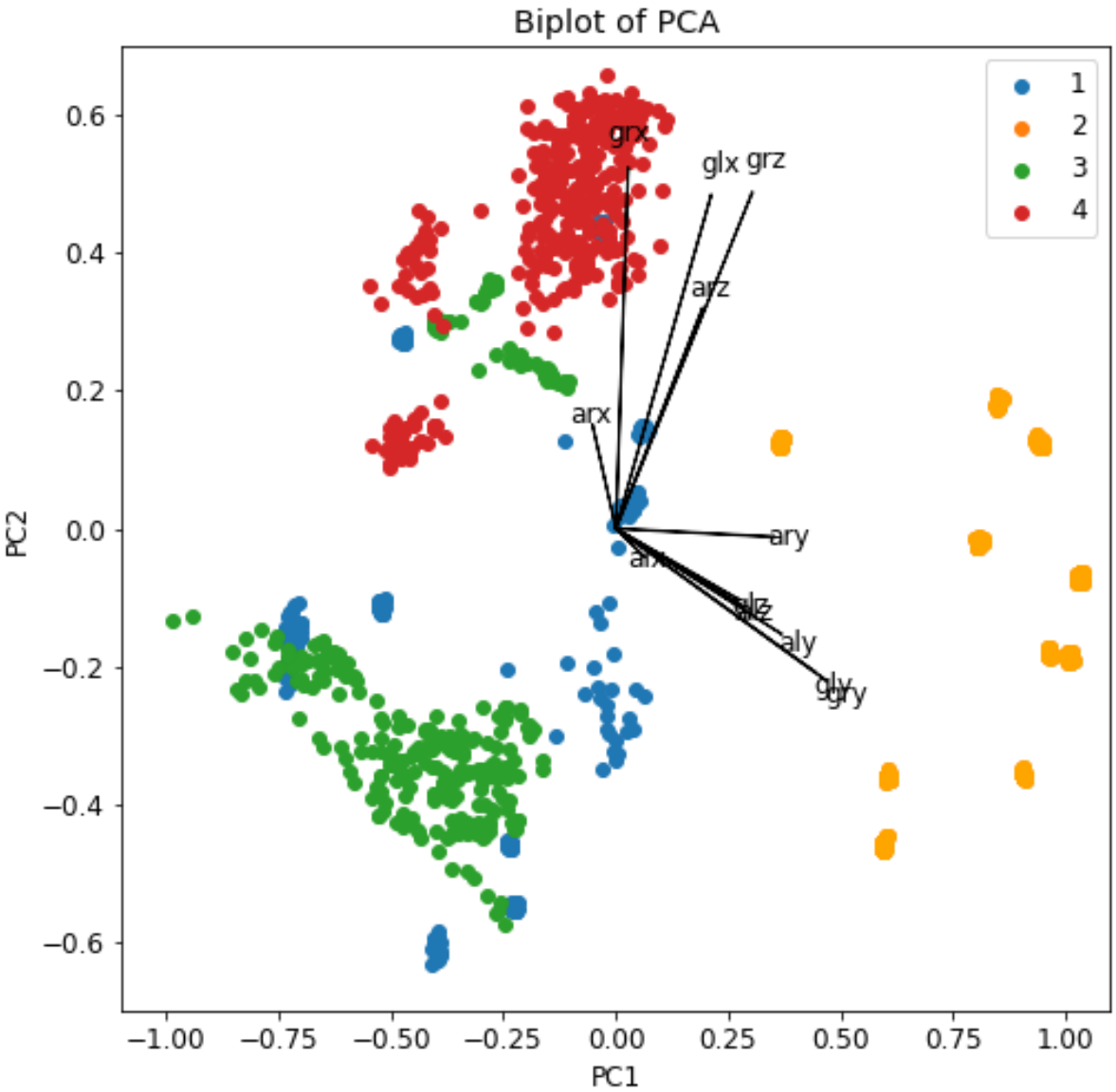


Q.3.3. Are there features that clearly separate class 2 from the other ones? [10 points]

□ Yes

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).

As can be seen down below class 2 (orange), is completely oriented in the positive side of PC1. class 2 is the only class with overly positive PC1 scores which differentiates it even more from the other classes. A feature is because of this not hard to find, generally for the entire class ary would fit well. Since it gives a positive PC1 score, but its effect on PC2 is insignificant which is similar to the datapoints of 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

Write some text and use a figure (see figure below the interpertation) to explain your reasoning (max 100 words).

Walking generally has a positive PC1 value with it's PC2 value being slightly positive or a bit negative. Jogging on the other hand mostly has a small to decently sized negative PC1 value and a positive PC2 value. This aligns better with the values from the gyroscopes then the data from the acceleration, the gyroscope data points towards a positive PC2 value with slightly positive PC1 value and neutral or positive PC1 value with a slightly negative PC2 value. The gyroscopic data also has more influence than the acceleration data as can be seen in the length of the arrows.

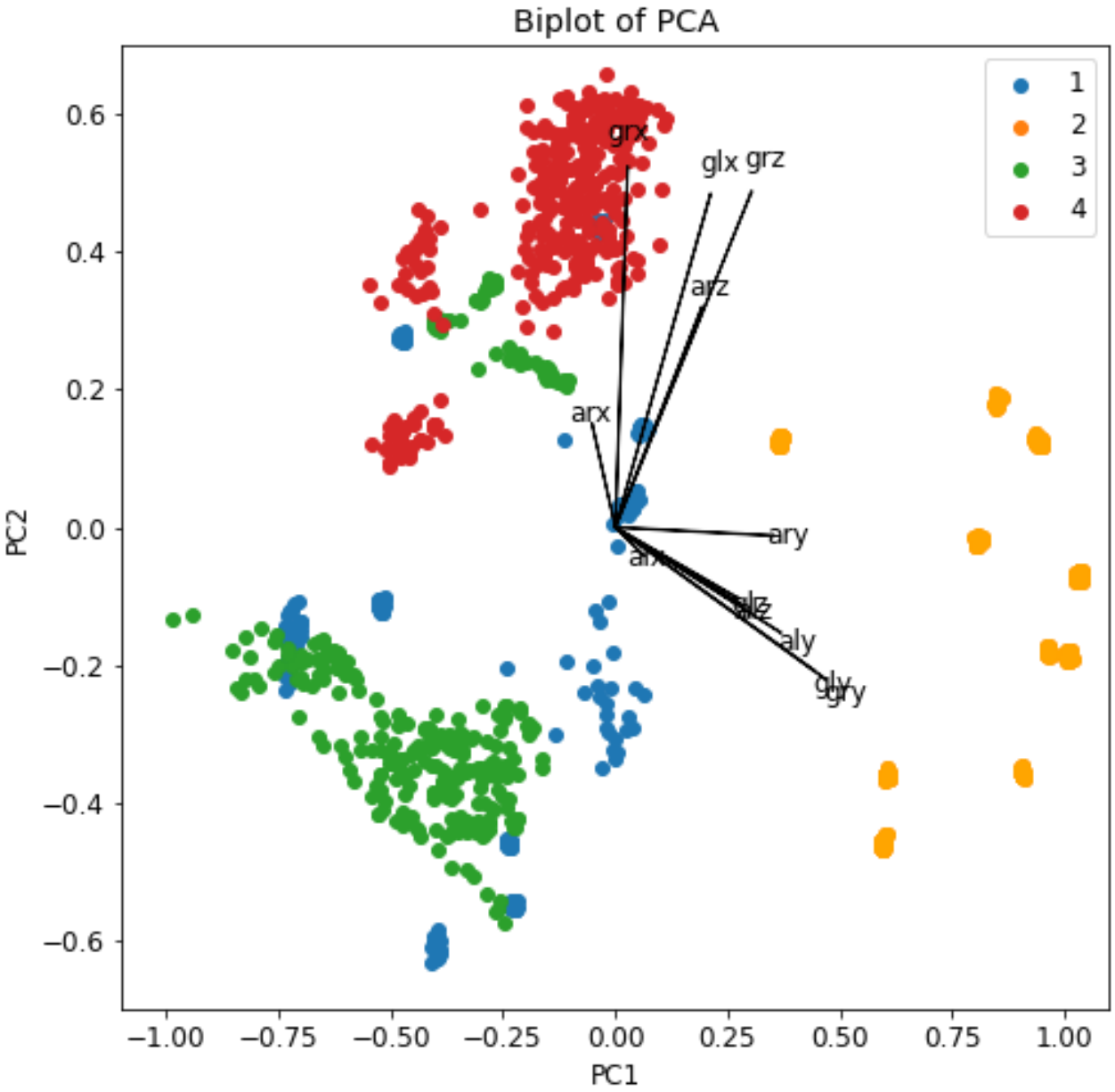
If you had to attempt an interpretation of what you just noted with your data analysis, what would it be? (max 300 words) [bonus question, it adds up to 10 extra points if your answer is interesting and well-thought of]

As can be seen there are clear clustered for different activities, from this we can find out which an activity is more influenced by PC1 or PC2. Activity 4 for example is clearly positively influenced by PC2, but only slightly by PC1. Furthermore activity 2 is very influenced by PC1 but less so by PC2. Activity 2 is also very scattered, but a lot of points do overlap with one another. Then there's activity 1 and 3 who both are mostly negatively effected by both pc2 and PC1. For activity 3 counts is more negatively effected by PC1 than for PC2, but for activity 1 this seems to be the opposite.

Activity 1 in general is pretty interesting, there's a lot of point concentrated around the 0,0 point and there's a lot of distance between small clustered of activity 1. Activity 3 also has this outlier cluster that gets positively influenced by PC2.

Another interesting thing about the figure is that the loading arrows don't really align well with any activity except for activity 2. Activity 1 and 3 don't align at all with the loading of the sensors, activity 4 meanwhile aligns a bit especially with grx, but not really with any other loadings.

Another interesting observation about the loadings is that the gyroscopic data has two groups that points in similar directions. Grx, glx and grz all point mainly toward a positive PC2 and slightly positive PC1 and the other group of gry, gly and glz, all point toward a positive PC1 and negative PC2 value. The acceleration in contrast points all over the place without any cohesion really, but negative PC1 values are still mostly avoided.


Biplot of PCA
1
2
3
4
grx
glx grz
arz
arx
ary
aly
gly
gry
PC1
PC2
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6