**Assignment 2**

Martijn Stoorvogel

**Guidelines**

This template contains the questions about assignment 2 and will need to be filled out with your answers. Please follow the instructions on the maximum number of words and figures allowed.

General guidelines:

* Make your figures standalone with figure titles, axis titles, legends, and captions.
* A figure containing multiple subplots counts as a single figure. However, use only the figures you need for your explanation, not more, not fewer.
* If you can reuse a figure to answer to multiple questions, you can consider numbering it and referencing it where appropriate.
* Write clearly and concisely. Answers that are not easy to understand or ambiguous will not be counted as correct.

Once you are finished with your answers, you can remove the blue text from this document, and export it as a pdf for submission. Rename your file using your initials and surname, e.g.: “FGrisoni\_PCA\_assignment.pdf”.

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

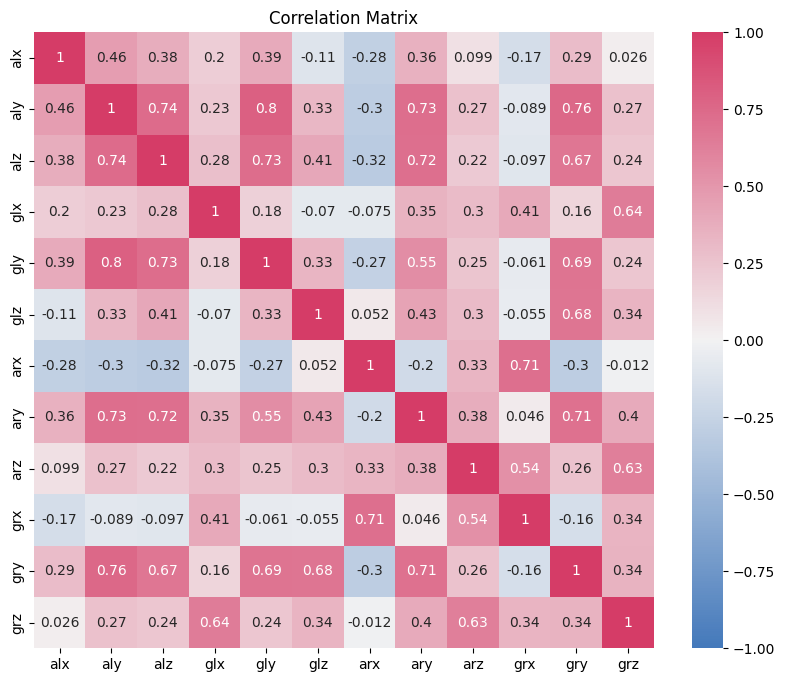
In this part of the assignment, you will reflect on what is contained in your data and how you can pretreat your data.

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]

**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) | 1. Different variable scales: If the variables in your dataset are measured in different units or have different scales, scaling is generally recommended. 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.  Scaling the variables before PCA ensures that each variable contributes proportionally to the analysis based on their variances. |
| *Unitary variance* (variance equal to 1) | By applying unit variance scaling, each variable will have a variance of one, making them comparable in terms of their variances. This allows PCA to focus on the relative importance of the variables based on their covariance structure rather than their individual variances. This scaling method focuses solely on adjusting the variances of the variables to be equal to one while preserving their means. |
| *MinMax scaling* (minimum and maximum equal to 0 and 1, respectively) | 1. Preservation of original scale: Min-max scaling transforms the variables to a specific range, typically [0, 1]. If preserving the original scale and units of the variables is important in your analysis, then min-max scaling may not be suitable. Min-max scaling can distort the original units of the variables, making it difficult to interpret the results in the context of the original data. 2. Sensitivity to outliers: Min-max scaling is sensitive to outliers, as it is influenced by the minimum and maximum values of the variables. Outliers can disproportionately affect the scaling, particularly when the range is small. If your data contains outliers and you are concerned about their influence on the PCA analysis, other scaling methods, such as standardization (Z-score normalization), may be more robust. 3. Equal treatment of variables: Min-max scaling ensures that all variables have the same range, which can be useful if you want each variable to contribute equally to the PCA analysis. However, keep in mind that variables with larger ranges may still dominate the analysis due to their larger spread, even after scaling. 4. Non-linear relationships: Min-max scaling assumes a linear relationship between the variables and their scaled counterparts. If the relationships are non-linear, min-max scaling may not capture the underlying patterns accurately. In such cases, non-linear scaling techniques like normalization by rank or power transformations might be more appropriate.   In summary, while min-max scaling can be applied to PCA analysis, it may not always be the best choice depending on the specific characteristics of your data. It's important to consider the original scale and units of the variables, the presence of outliers, and the linearity of the relationships before deciding on the appropriate scaling method for your PCA analysis.  Bovenkant formulier |
| *Standard scaling* (mean equal to 0 and variance equal to 1) | 1. Preservation of original scale: Standard scaling transforms the variables to have a mean of zero and a standard deviation of one. This standardized form may deviate significantly from the original scale and units of the variables. If preserving the original scale and units is important for your analysis and interpretation, standard scaling may not be the most suitable choice. 2. Sensitivity to outliers: Standard scaling is influenced by the mean and standard deviation of the variables. It is more robust to outliers compared to min-max scaling since it uses statistical measures that are less affected by extreme values. If your data contains outliers and you are concerned about their impact on the PCA analysis, standard scaling may be a more appropriate choice compared to min-max scaling. 3. Equal treatment of variables: Standard scaling ensures that each variable has a mean of zero and a standard deviation of one. This equalizes the variances across variables, allowing PCA to focus on the relative importance of the variables based on their standard deviations. Variables with larger variances will contribute more to the principal components than those with smaller variances. 4. Linearity assumption: Standard scaling assumes a linear relationship between the variables and their standardized counterparts. If the relationships are nonlinear, standard scaling may not accurately capture the underlying patterns. In such cases, alternative scaling methods or non-linear transformations may be more suitable. 5. Gaussian distribution assumption: Standard scaling assumes that the variables are approximately normally distributed. If the variables in your data significantly deviate from a normal distribution, the effectiveness of standard scaling may be compromised. In such cases, it may be beneficial to explore other scaling techniques that are more robust to non-normality. |

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

[figure here]

**Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving**

**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

\_\_\_\_\_\_

Visualize your answers with a figure and report it below.

[figure here]

Afbeelding met tekst, lijn, Perceel, diagram

Automatisch gegenereerde beschrijving

*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).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

grx

\_\_\_\_\_\_

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

* Yes
* No

Motivate your answer (max 50 words)

gry

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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

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

[figure here]

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]

* 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]

**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).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[figure here]

**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).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[figure here]

**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).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[figure here]

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]