Sigurður Ágúst Jakobsson

# Project: Assignment 12 – Principal Component Analysis

## Section 1.2

My results below. Same shape but slightly different scale. I cant see why, the standardizing code is very simple and I fail to see my error.

Chart, scatter chart

Description automatically generated

## Section 1.3

Diagram

Description automatically generated

## Section 1.4

Judging from all these plots, which dimensions do you think correlate best with the first dimension?

Using the information [here](https://file+.vscode-resource.vscode-cdn.net/c%3A/Users/Siggi/Desktop/Sk%C3%B3linn/%C3%96nn%209/Gagnan%C3%A1m%20og%20vitv%C3%A9lar/Programming%20Assignments/T809DATA_2022/08_SVM/README.md), what are the names of those features?

Can you explain why it makes sense that these are the dimensions that best correlate with the first one as opposed to the other dimensions?

The key for the cancer data from assignment 8 is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| [0]radius\_mean | [1]texture\_mean | [2]perimeter\_mean | [3]area\_mean | [4]smoothness\_mean | [5]compactness\_mean |
| [6]concavity\_mean | [7]concave points\_mean | [8]symmetry\_mean | [9]fractal\_dimension\_mean | [10]radius\_se | [11]texture\_se |
| [12]perimeter\_se | [13]area\_se | [14]smoothness\_se | [15]compactness\_se | [16]concavity\_se | [17]concave points\_se |
| [18]symmetry\_se | [19]fractal\_dimension\_se | [20]radius\_worst | [21]texture\_worst | [22]perimeter\_worst | [23]area\_worst |
| [24]smoothness\_worst | [25]compactness\_worst | [26]concavity\_worst | [27]concave points\_worst | [28]symmetry\_worst | [29]fractal\_dimension\_worst |

Dimension 1 (index 0) of course correlates well to itself on the first plot. The other two visually clearest correlations are dimensions 3 and 4 (indexes 2 and 3). These are [2]perimeter\_mean and [3]area\_mean, that correlate to [0]radius\_mean. I think it makes sense that they correlate since they are all geometric measurements. The radius has an effect on the perimeter (circumference) and area. Circumference = 2 \* Pi \* radius and Area = Pi \* radius ^2. The others are a lot more about the physical characteristics, standard deviations and irregularities and don’t have these clear geometric relations.

## Section 2.1

Diagram, engineering drawing

Description automatically generated

## Section 3.1

Chart, line chart

Description automatically generated

## Section 3.2

Chart, line chart

Description automatically generated

## Section 3.3

## Chart, line chart Description automatically generated

## Section 3.4

## Using your knowledge about PCA and your plots, can you explain why they look like they do? Explain why the eiginvalue plot has the trend that it does and the same for the cumulative variance plot.

Eigenvalues in PCA explain the amount of variance that a component represents. The First eigenvalues are often much larger than others and we choose the first M of these to maintain variance in the transformation. This is why the eigenvalue plot and the log eigenvalue plot decrease. The cumulative variance plot shows how much of the overall variance is represented by the first M components and therefore it rises fast since the first components are so large compared to subsequent ones.

## Independent Section

Make an experiment of your choice, relevant to PCA, and present your results and all added code with plots and figures is applicable.