Homework 1: Principal Component Analysis (PCA)

Theresa Wakonig
Data Mining II

March 16, 2021

1 Exercise 1

Proof of trace maximization approach.

Let $U \in \mathbb{R}^{d\times r}$ and $\vec{x}_i \in \mathbb{R}^d$.

$argmin \sum_{i=1}^{n} ||\vec{x}_i - UU^T\vec{x}_i||_2^2 = argmin \sum_{i=1}^{n} (\vec{x}_i - UU^T\vec{x}_i)^T(\vec{x}_i - UU^T\vec{x}_i)$

$$\begin{aligned}
(\vec{x}_i - UU^T\vec{x}_i)^T(\vec{x}_i - UU^T\vec{x}_i) &=^{i)} trace((\vec{x}_i - UU^T\vec{x}_i)(\vec{x}_i - UU^T\vec{x}_i)^T) \\
&= trace((\vec{x}_i - UU^T\vec{x}_i)(\vec{x}_i^T - (UU^T\vec{x}_i)^T)) \\
&= trace((\vec{x}_i - UU^T\vec{x}_i)(\vec{x}_i^T - \vec{x}_i^T UU^T)) \\
&= trace(\vec{x}_i\vec{x}_i^T - 2\vec{x}_i^T UU^T\vec{x}_i + \vec{x}_i^T U \underbrace{U^T U}_{\mathbb{I}} U^T\vec{x}_i) \\
&= trace(\vec{x}_i\vec{x}_i^T - \vec{x}_i^T UU^T\vec{x}_i) \\
&=^{ii)} trace(\vec{x}_i\vec{x}_i^T) - trace(\vec{x}_i^T UU^T\vec{x}_i) \\
&= \underbrace{||\vec{x}_i||^2}_{const.} - trace(U^T\vec{x}_i\vec{x}_i^T U)
\end{aligned}$$

$\Rightarrow$ constants can be left out
$\Rightarrow$ minimizing the negative trace is equal to maximizing the positive trace

$$argmax \quad trace(U^T \sum_{i=1}^{n} \vec{x}_i\vec{x}_i^T U)$$

2 Exercise 2

Implementation of PCA.

2.a

See *pca.py*.

2.b

See *pca.py*.

2.c

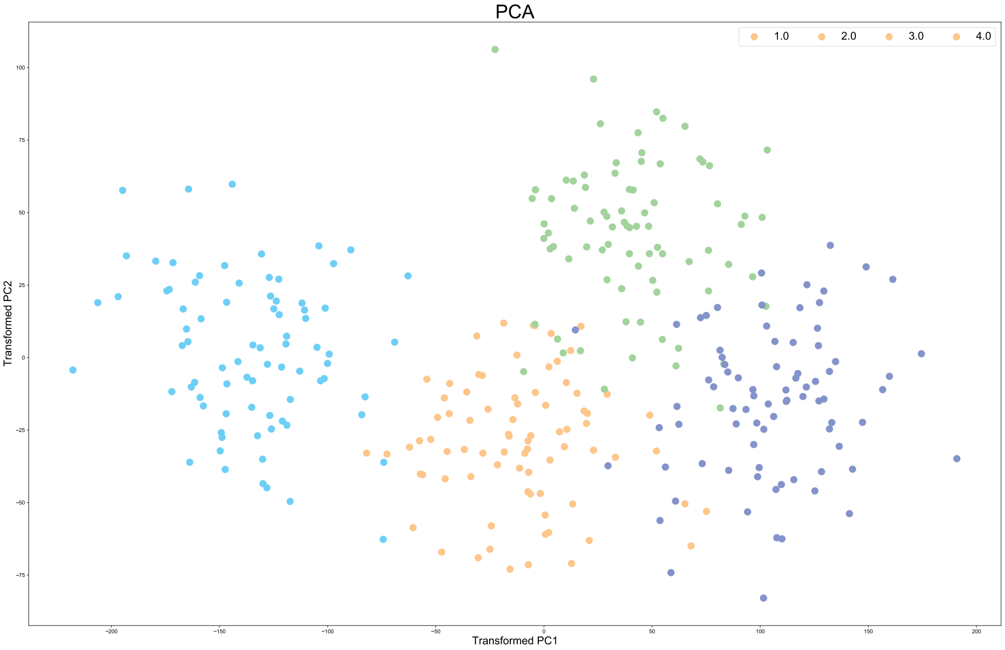


Figure 1: Transformation of the input data onto two-dimensional subspace with PCA. The color code in the legend should be as follows: 1.0 - yellow, 2.0 - green, 3.0 - light blue, 4.0 - purple.

By using the first two principal components of the data, the dimensionality was reduced from 40 to 2. The colors in the plot refer to the class labels of the observations. Overall, the transformed data depicts four clusters/groups which are not clearly separated. Slight overlap of clusters can be observed. Without the color code the clusters would be indistinguishable from one another. Moreover, the scatter plot illustrates more variation along the PC1-axis than along the PC2-axis. This was to be expected from our definition of the principal components discussed in the lecture.

2.d

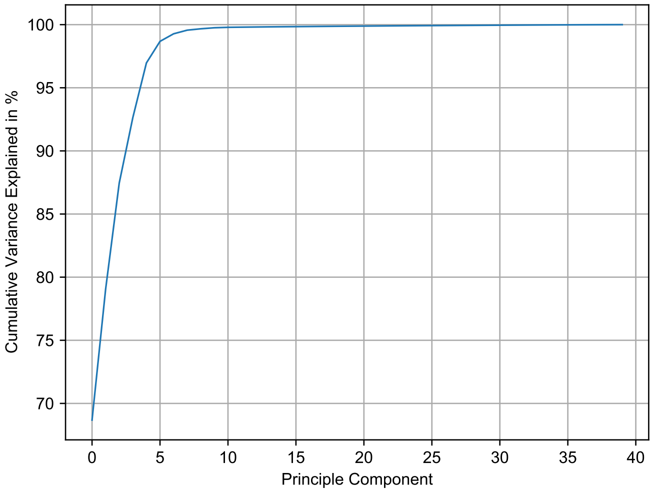
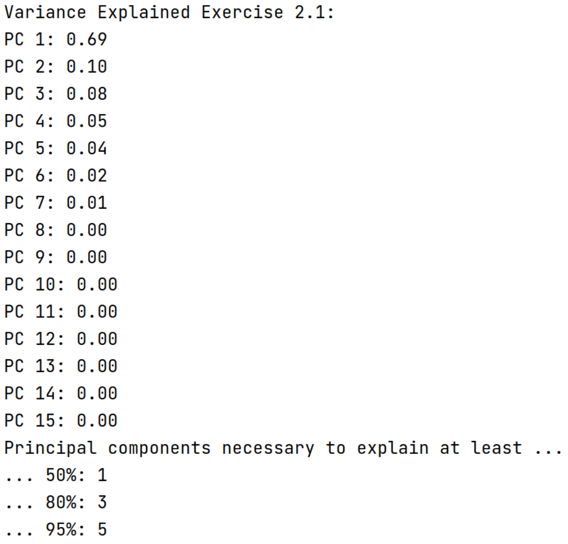


Figure 2: Variance in the data (in %) that can be explained by the principal components.

The graph depicts the percentage of variance in the data that can be explained by taking the respective number of principal components into account. From our definition of the principal components, the first principal component defines the direction in the data with the greatest variance while the last PC defines the direction with the least variance in the data.
This dependence is in general not linear, as can also be seen in the graph above. The rapid convergence and steep slope of the graph lead to the conclusion that we only need the first three to five PCs to explain the total variation in data. As discussed in 2e) one has to be careful when making such assumptions. The corresponding program output can be seen in the screenshot below.



```
Variance Explained Exercise 2.1:
PC 1: 0.69
PC 2: 0.10
PC 3: 0.08
PC 4: 0.05
PC 5: 0.04
PC 6: 0.02
PC 7: 0.01
PC 8: 0.00
PC 9: 0.00
PC 10: 0.00
PC 11: 0.00
PC 12: 0.00
PC 13: 0.00
PC 14: 0.00
PC 15: 0.00
Principal components necessary to explain at least ...
... 50%: 1
... 80%: 3
... 95%: 5
```

2.e

Results for exercise 2.2 with normalized data:

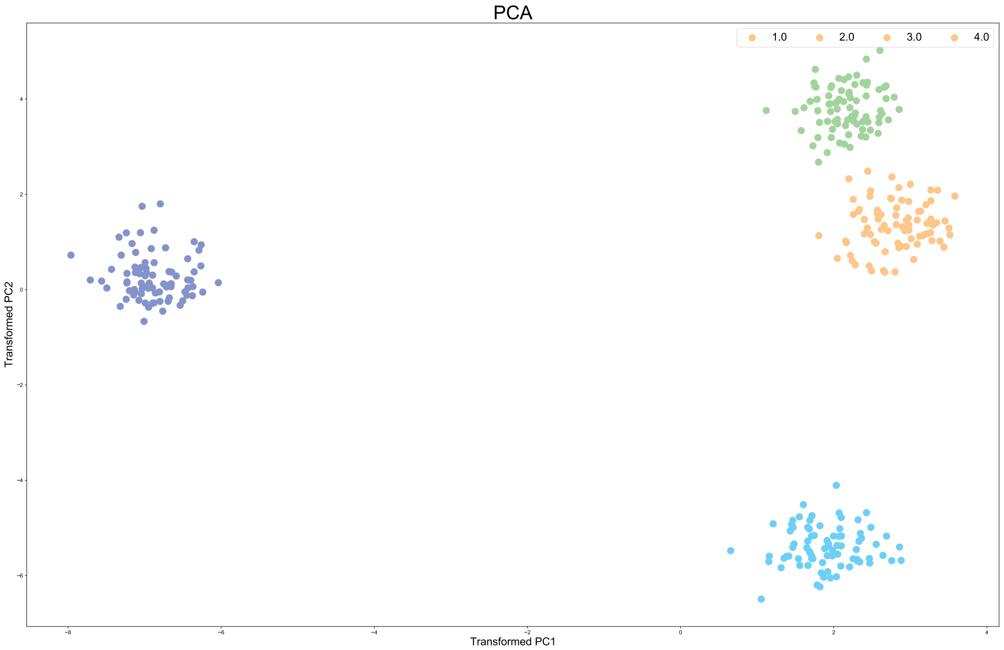


Figure 3: Transformation of the input data onto two-dimensional subspace with PCA. The color code in the legend should be as follows: 1.0 - yellow, 2.0 - green, 3.0 - light blue, 4.0 - purple.

The scatter plot depicts four clusters which are clearly separated from each other. Without knowing the class labels of the observations in advance, one would very nicely be able to put the points into four groups. The PCA clearly helped to find the underlying trends in our data.

2.f

The graph converges a lot slower when using normalized data. This means that a single PC does not have as much weight as previously seen in 2d).

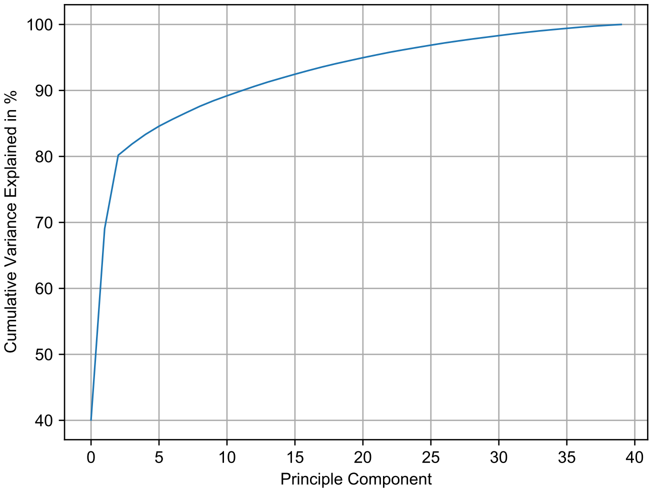


Figure 4: Variance in the data (in %) that can be explained by the principal components.

```
Variance Explained Exercise 2.2:
PC 1: 0.40
PC 2: 0.29
PC 3: 0.11
PC 4: 0.02
PC 5: 0.01
PC 6: 0.01
PC 7: 0.01
PC 8: 0.01
PC 9: 0.01
PC 10: 0.01
PC 11: 0.01
PC 12: 0.01
PC 13: 0.01
PC 14: 0.01
PC 15: 0.01
Principal components necessary to explain at least ...
... 50%: 2
... 80%: 3
... 95%: 22
```

Normalization (centering and dividing by std. deviation) ensures that the data follows a $\mathcal{N}(0,1)$ distribution. This is necessary in order to get rid of units. As a variable with a high standard deviation (which may be solely caused by the choice of units) shows high variance, leaving out the normalization will in many applications naturally lead to the behaviour we witnessed in the first part of this exercise: Very few underlying trends explain almost all the variance in the data. By normalizing we ensure that all variables are given the same ’weight’ in our transformation.