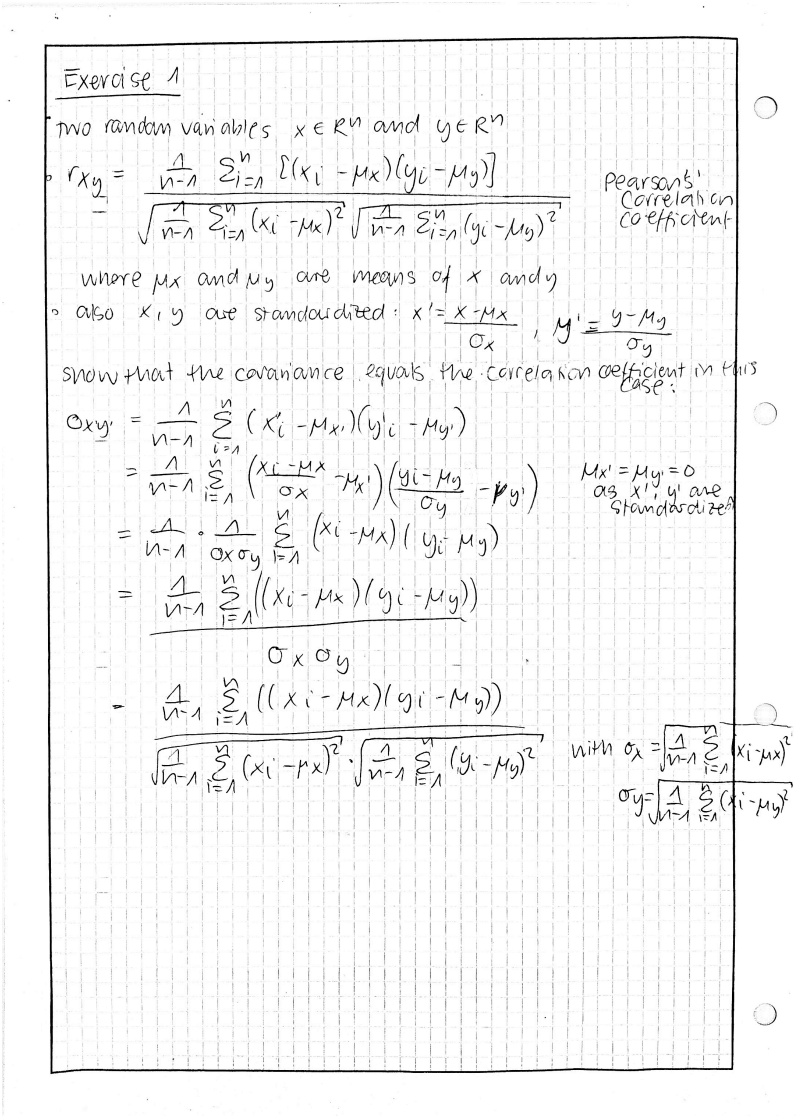
Data Mining 2 Homework 2

Jennifer Probst

31 March 2021

Exercise 1: Proof



Exercise 1

two random variables $x \in R^n$ and $y \in R^n$

- $r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i-\mu_x)(y_i-\mu_y)]}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu_y)^2}}$ Pearson's Correlation coefficient

where μ_x and μ_y are means of x and y

- also x, y are standardized: $x' = \frac{x-\mu_x}{\sigma_x}$, $y' = \frac{y-\mu_y}{\sigma_y}$

show that the covariance equals the correlation coefficient in this case:

$$\sigma_{x'y'} = \frac{1}{n-1}\sum_{i=1}^{n}(x'_i-\mu_{x'})(y'_i-\mu_{y'})$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}-\mu_{x'}\right)\left(\frac{y_i-\mu_y}{\sigma_y}-\mu_{y'}\right) \quad \mu_{x'}=\mu_{y'}=0 \text{ as } x', y' \text{ are standardized}$$

$$= \frac{1}{n-1}\cdot\frac{1}{\sigma_x\sigma_y}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}((x_i-\mu_x)(y_i-\mu_y))}{\sigma_x\sigma_y}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}((x_i-\mu_x)(y_i-\mu_y))}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)^2}\cdot\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu_y)^2}}$$

with $\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)^2}$, $\sigma_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_y)^2}$

Exercise 2: PCA / SVD comparison

a) The following Figure 1 show the results of an PCA dimensionality reduction to a 2-dimensional subspace of their first two principal components. In the plot the classes are colored by their class label (three different flower classes). Furthermore Figure 2 shows the cumulative variance explained for each of the four principle components. They explain 0.92 percent of the variance (first PC), 0.05 percent (second PC), 0.02 percent (third PC) and 0.01 percent in case of the fourth PC. The first PCs explain almost all the variance, whereas the other PCs don't contribute much to the variance explained.

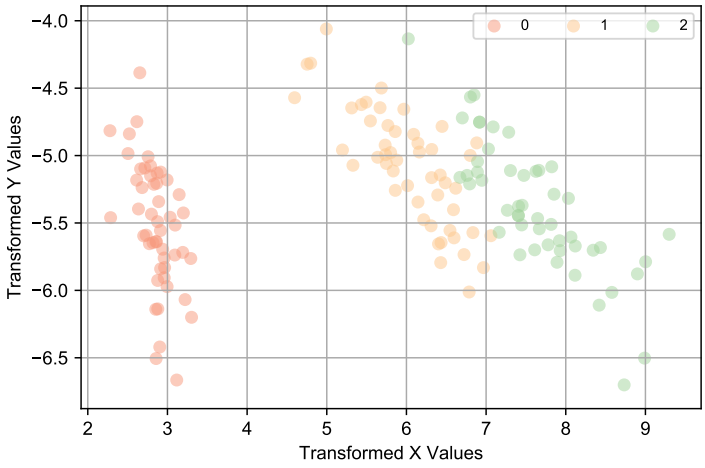


Figure 1: Plot of the transformed data in their first two PCs. Samples are colored according to their class labels.

b) This is a pseudocode for a PCA using SVD: We require a data matrix D with dimensions dxn, where d is the dimension of the n samples. Let r be the number of dimensions the data should be reduced to.

- First we mean-center the data matrix D. Therefore the from each sample value the mean of this sample gets subtracted and divided by the standard deviation of this sample.
- In a next step if n>d, we compute the singular value decomposition of D which is a product of L, ss.T and L.T. This decomposition coincides with a spectral decomposition of D*D.T (matrix product of the matrix D and its transpose) meaning L represents the eigenvectors of D*D.T and ss.T the eigenvalues of D*D.T. As D*DT / n is the covariance matrix of mean centered data D, we find that the eigenvalues ss.T / n are the same than the ones obtained by PCA and the EV of PCA are the same as left singular vectors of SVD (in matrix L) for mean centered data.
- We now sort the eigenvectors according to their eigenvalue as done of PCA as well. These can now be used to project the data to a lower dimensional representation. If n<d we the singular value decomposition to get the spectral decomposition of D.T*D.

c) see code

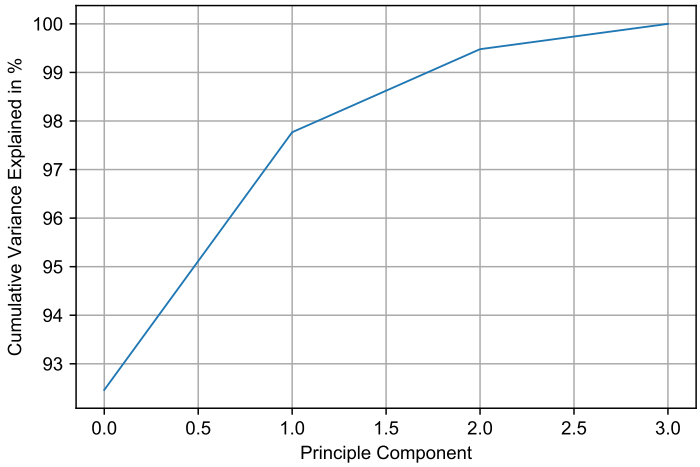


Figure 2: Plot of the cumulative variance explained by all principal components.

d) The following Figure 3 show the results of an dimensionality reduction to a 2-dimensional subspace using SVD (SVD based PCA). In the plot the classes are colored by their class label. We see that the plot is exactly the same as the one for PCA dimensionality reduction (Figure 1), the data is only mirrored. This might be as the case as the PCs coincide value wise with the left singular vectors, but their sign might be opposite. Furthermore Figure 4 shows the cumulative variance explained by the four left singular vectors (LSV). They explain 0.92 percent of the variance (first LSV), 0.05 percent (second LSV), 0.02 percent (third LSV) and 0.01 percent in case of the fourth LSV. We find that these numbers are exactly the same as in Figure 2 as the PCs coincide with the LSVs. Therefore we show that the principal component analysis based on SVD and PCA on covariance matrix result in same results.

e) The singular value decomposition should be much quicker. To obtain the spectral decomposition in PCA we have to compute an inverse, which is often costly in runtime. The SVD might therefore be used to obtain eigenvalues and eigenvectors much quicker.

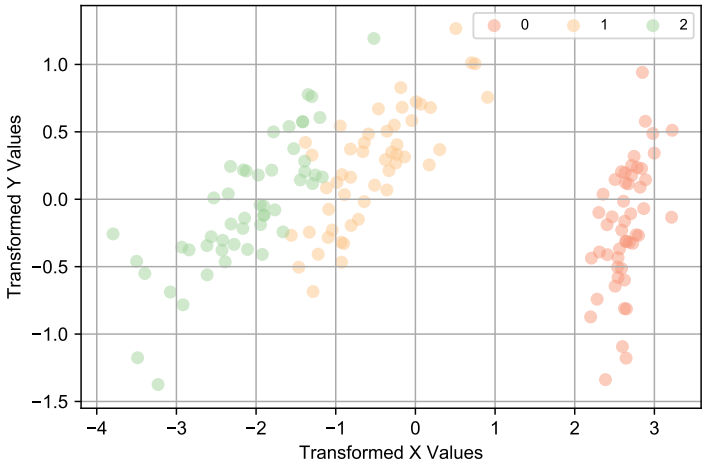


Figure 3: Plot of the transformed data using the two left singular vectors with highest eigenvalue from SVD. Samples are colored according to their class labels.

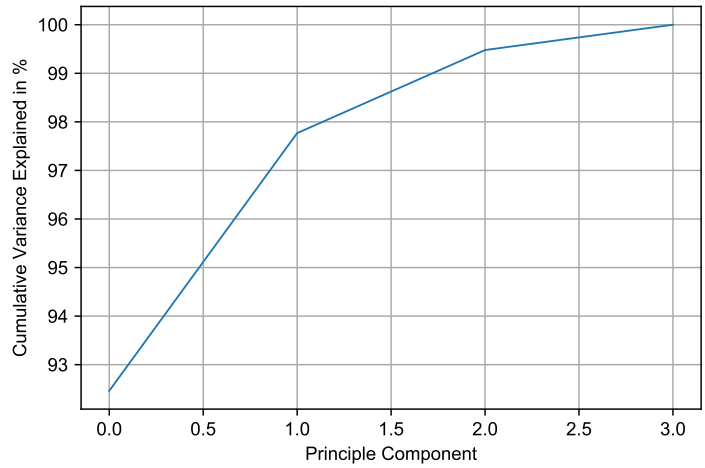


Figure 4: Plot of cumulative variance explained by all left singular values.

Exercise 3: Moore-Penrose pseudo-inverse

see code / We observe that both equations hold.