Medical Image Processing - Assignment 1

Group 13

Emirhan Kurtulus
12243493
Nikolaus Czernin
11721138
Hossein Monsefi Estakhrposhti
12231205

ABSTRACT

In this assignment, we investigate covariance, Principal Component Analysis (PCA), and shape modeling using 2D and 3D datasets. For the first task we analyze covariance matrices to understand relationships between features, then we apply the PCA to identify directions of variance and perform dimensionality reduction. In the projection and reconstruction phase we will try to show how variance is kept or changed according to the principal components used. Lastly, we use the PCA-based shape modeling to generate realistic bone shapes, by illustrating how variance affects shape accuracy and smoothness.

KEYWORDS

PCA, Covariance, Eigenvalues, Dimensionality Reduction, Data Projection, Shape Modeling

1 TASK 1: COVARIANCE MATRICES

We have three 2-dimensional datasets that we are using in this part of the exercise we also have one 3-D dataset that we do not need to use for this part of the exercise, and therefore their covariance matrices are square matrices with 2 rows and columns. The values on the top-left to bottom-right diagonal are the variances of the attributes, i.e. the first diagonal value is the spread of the dataset's first attribute etc.. The remaining diagonal values, which are mirrored along the diagonal, are the covariances of the different attributes, i.e. the linear relationship of the two attributes, whereas a value close to 1 symbolizes a strong linear relationship and thus a tendency to form a bottom-left to top-right diagonal orientation on the scatterplot. From the figure 1 we can see that in the first dataset, in which the two variables clearly correlate positively, there is a positive covariance of 0.67673206. In the second dataset, on the other hand, their correlation is obviously negative, and their covariance is accordingly negative, -0.90833726. Additionally, we see that from the last figure, which is located on the top right, there is no covariance between the features.

$$\text{Covariance matrix for } \texttt{data1} = \begin{bmatrix} 0.9380 & 0.6767 \\ 0.6767 & 1.0356 \end{bmatrix}$$

$$\text{Covariance matrix for } \texttt{data2} = \begin{bmatrix} 0.9287 & -0.9083 \\ -0.9083 & 0.9914 \end{bmatrix}$$

$$\text{Covariance matrix for } \texttt{data3} = \begin{bmatrix} 0.8877 & 0.0167 \\ 0.0167 & 1.0060 \end{bmatrix}$$

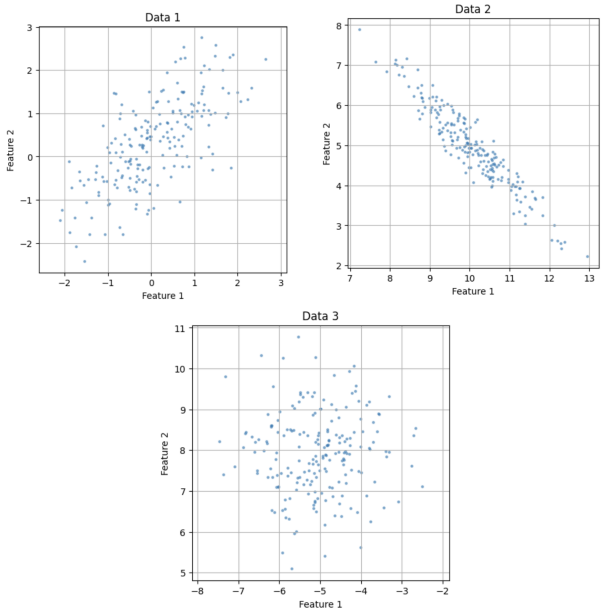

Figure 1: Scatterplots of data1 (top-left), data2 (top-right) and data3 (bottom)

2 TASK 2: PCA

2.1 What do the eigenvectors represent?

In the figure 2, we see the eigenvectors, *the red lines in the scatterplots*, represent *the variance of the attributes*, where the lines point in the direction in which the attributes vary the most and the length of the lines represent how much they vary in that direction.

2.2 What do the eigenvalues represent?

The eigenvalues quantify how much of the original attribute's variance can be explained by the corresponding eigenvector, where a *higher eigenvalue means more variance* explained. If there are some high eigenvalues and some very low ones, that means that of the projected data, keeping only the ones with a high corresponding eigenvalue would leave you with data that is of lower dimension but still explains most of the data's variance, which is enables dimensionality reduction.

2.3 How does omitting the mean subtraction from the data matrix X affect the computation of PCA?

The mean is like the centre of mass in the given data. PCA rotates the dataset to find maximum variances along the base-axes and then projects that rotation to the original data. But if the data does not have its center, i.e. its mean, at the origin, the first principal components would also adjust toward the mean and not just represent parts of the original variance. The impact of this should be larger for dataset 1 and 2 than for dataset 0, as dataset 0 is already almost mean-centered as is.

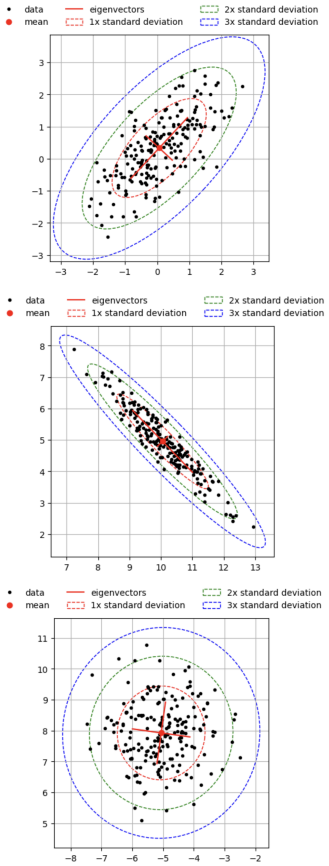

Figure 2: PCA-Scatterplots of data1 (left), data2 (center) and data3 (right)

3 TASK 3: SUBSPACE PROJECTION

3.1 PCA Projection and Reconstruction on 2D Data

For the first part of the task, we performed the PCA on the Data1, which has 2 dimensions. Then we projected the data onto the first principal component, and we plot the projected data. Moreover, projecting the data onto only the first principal component reduces the data dimensionality to 1. We have the shape as (200,1), which means the dimensionality is 1D.

Then we reconstructed the data from the projection back into the original 2D space. Next, we plotted the data by using the helper function. We see the reconstructed data points (green stars) lie along the first principal component axis. Compared to the original scatters as we see with the black points, the reconstructions are approximations and do not completely capture the variation to the first principal component. As we can clearly see in data1's plot, where the reconstructions follow the dominant trend, but better spread is lost.

From the table 1, we see that the comparisons of reconstruction error by using the mean squared error, for the data1, the error is 0.0509, which means the projection captures most of the variance. For the data2, we have the error as much higher at 0.8809, which shows poorer reconstruction.

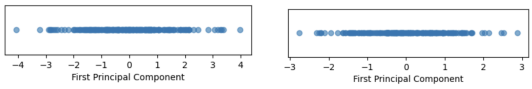

Figure 3: Projection onto first principal components of data1 (top), data2

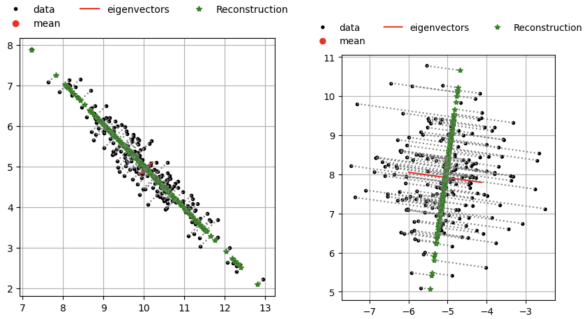

Figure 4: Reconstruction of first principal components into 2D of data1 (top), and data2

Dataset	MSE (1st PC)	MSE (2nd PC)
data1	0.0509	1.8596
data2	0.8809	1.0033

Table 1: Mean Squared Reconstruction Error after projection onto the 1st and 2nd principal components

For the following part of the task, first we repeated the steps from the previous part, but this time we projected the data onto the second principal component. For the data1 we see the projection is on the second PC, which captures little variance, for the same dataset we got the reconstruction error as 1.8596, which is much higher from the first PC (0.0509) for the same dataset. From the plot for the same dataset which can be seen in the figure **??**, green reconstructed points are close to the mean line, losing much of the

data's spread along the main direction. Additionally, we can say that the projection plot is very narrow, indicating minimal variance. When we look at the reconstruction error for the data, it has the error 1.0033, which is lower compared to data1, yet still performs badly compared to the first principal component for the same data. In general, data2 has more spread across both principal components, but the first PC captures more structure than the second PC.

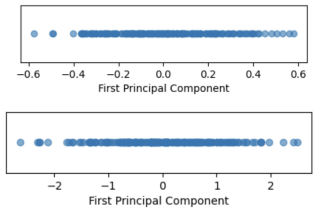

Figure 5: Projection onto second principal components of data1 (left), data2 (right)

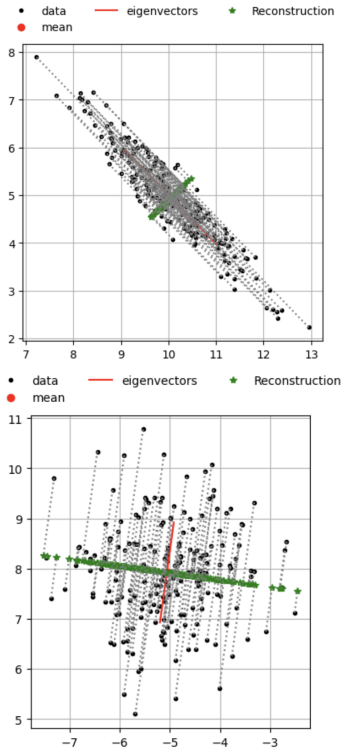

Figure 6: Reconstruction of second principal components into 2D of data1 (left), data2 (right)

To summarize, in the light of the information given before, it makes sense to choose the *first principal component for both datasets*. Based on the reconstruction error, we see that it is significantly lower when we use the first principal component.

4 TASK 4: INVESTIGATION IN 3D

4.1 PCA and 3D Visualization

For the first phase of this task, we performed the PCA on the given 3-dimensional data. From the figure 7 we see that the first principal component, which is represented by green, likely correspond to the largest eigenvalue. Furthermore, the ellipsoid shape aligns with the eigenvectors, confirming that PCA has captured the data's main variation directions.

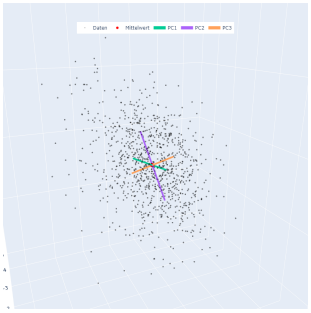

Figure 7: PCA of 3D Data

$$\begin{bmatrix} 1.03 & 0.79 & 0.54 \\ 0.79 & 1.45 & 0.34 \\ 0.54 & 0.34 & 1.01 \end{bmatrix}$$

In the matrix, which encodes how much each pair of dimensions vary together. Diangonal entries are the variances along the three dimensions(x,y,z), and off-diagonal values are correlations between axes.
Eigenvectors, they define the directions of maximum variance, they are the new orthogonal axes.
Eigenvalues are the amount of the variance that captured along each principal component. When it gets larger, we have more spread, or importance with another saying, that PC has.
When we consider the ellipsoid of standard deviations, we can say that it correspond to the axes of the ellipsoid that approximates the data distribution. With a more elongated axis we have higher variance in that direction. With a less, the other way around.

4.2 Projection and Reconstruction in Subspace

First we confirm that the dimensionality after projection is 2, as we see in the output as also shown below:

```
Shape of projected data3d: (1000, 2)
```

From the figure 11 below, we see that since the data was projected onto only two of three principal components, the variations along the third principal component was a bit lost. When we project the 3D data onto the first two principal components, the third principal component is discarded, due to that the results in a loss of variance and structure through that direction.
We see from the figure 11 that the reconstructed data points lie entirely within the 2D subspace spanned by PC1 and PC2. Additionally, we also see that variability that originally existed along PC3 is removed, which is leading to a flattening of the data in the

3D space. To show this lost clearly, we have the result of average reconstruction error for data3d as 0.3198, which illustrates that some portion of the dataset's structure has been lost due to the omission of PC3.
To summarize, it is clear to see that any features, structural design aligned with the third component, even-though it is small, were lost.

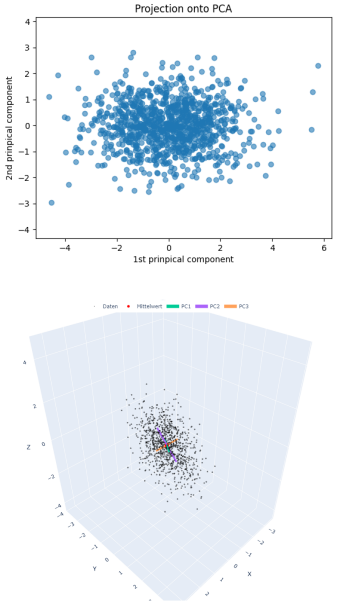

Figure 8: Projection onto PCA (left) 3-D presentation PCs

```
Average reconstruction error for data3d: 0.3198
```

5 SHAPE MODELING

In figure 9 we overlaid all bones' shapes as scatterplots.

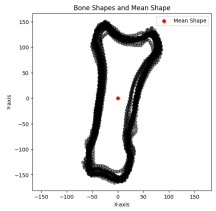

Figure 9: Shape of all the bones

In figure 10 we plotted the mean bone shape, reshaped to 2D, so 1 flattened shape per row, and overlaid a generated bone-shape from our created function, which perturbs the mean along the first two PCA modes. The generated shape (blue) is barely visible, as it closely matches the mean shape.

In figure 11 we visualize the geometric effects that the modes have on the shapes of new bone shape generations. The red points are, as before, the mean shapes, and the blue shapes are the modes' coefficients +- the eigenvalues of the PCA. Mode 0 for example

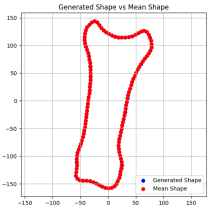

Figure 10: Shape modeling with PCA

represents the biggest variance patterns, forming a blue "cloud" around the mean bone shape and thus capturing the bone shape in the most general sense. The later modes, mode 10 for example, is then closer to the mean shape already, which features only very little variance, rather fine noise.

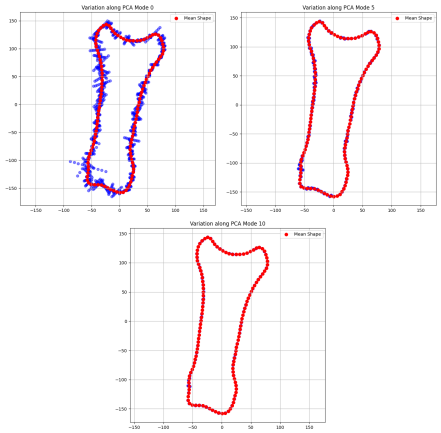

Figure 11: Shape modeling for different modes

In figure 12 we visualize the results of 4 different random bone generations, with varying variance threshold the minimum number of principal components required to reach the given variance level. Lowering the variance threshold appears to lead to more outlying points, which negatively affect the smoothness and thus the realism of the generated shapes. Matching less variance leads to a smoother bone shape, as the generation tries to match even less of the generated sample's variance to the ground truth distributions, which eventually leads to shapes that no longer match the distinct visual features of bones.

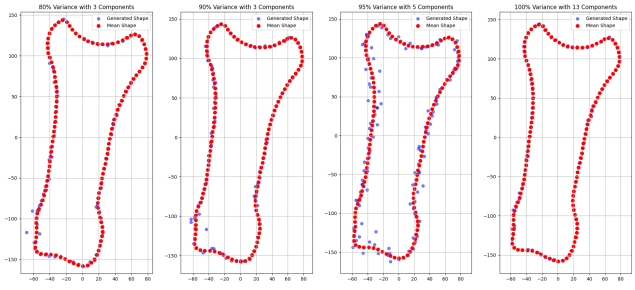

Figure 12: Comparisons for different thresholds