Miniproject 1

Kai Ye ye000064@umn.edu

December/16/2020

1 Low-dimensional Structure in High-dimensional Data

Experiment with the script using different values of the rank parameter r. To help aid your visualization, use the Rotate 3D feature of the MATLAB plotter to view the point clouds from different orientations.

1. Submit one representative scatter plot for each of the three cases: r = 1, r = 2, and r = 3, for the default case in the MATLAB script, where each of the (nonzero) eigenvalues of the covariance matrices is equal to one.

Solution:

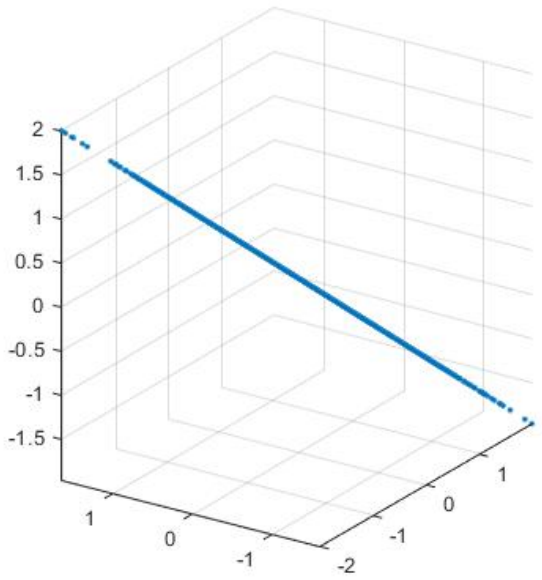

Figure 1: Scatter plot for $r = 1$

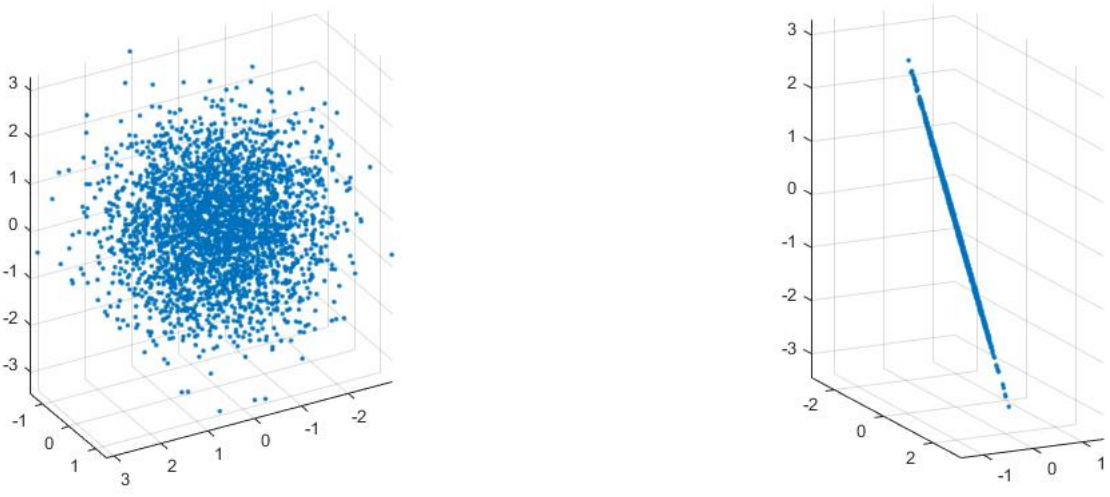

Figure 2: Scatter plot for $r = 2$ from two different views

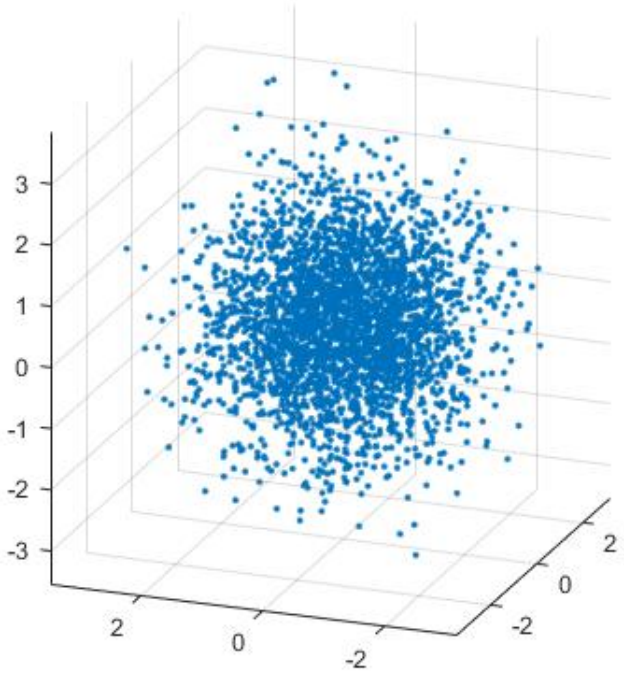

Figure 3: Scatter plot for $r = 3$

2. In the above experiments, what structure do you notice in the cases r=1 and r=2? More specifically, in each case the data vectors are all points in $\Re^3$ (by construction), but what can you say about the intrinsic structure or effective dimension of the data? You may need to rotate the point clouds to see this structure. Submit a written explanation of your observations.

Solution: The data vectors in the case $r = 1$ are generally in a 1-D line. The structure with $r = 2$ is generally a 2-D plane.

3. Now, experiment with the case where the (nonzero) eigenvalues of the covariance matrix are not all equal to one, but rather are generated randomly (as realizations of a uniform[0, 1] continuous random variable). How do these unequal eigenvalues affect the shape of the point cloud (especially when r = 2 and r = 3)? Submit a written explanation of your observations.

Solution: Compared to previous plots in q1, the shape of the point cloud is more condensed along certain direction as the cases when $r = 2$ and $r = 3$. Though the structures are still a plane and a 3-D body, the structures in part 1 are more like a circle and a sphere centered at the origin, while in this question they look like an ellipse and an ellipsoid and are shrunk .

The reason is the nonzero eigenvalues with that uniform distribution changes the covariance matrix, and make it more condensed along certain axis compared to constant ones, which indicates equal probability along all axises.

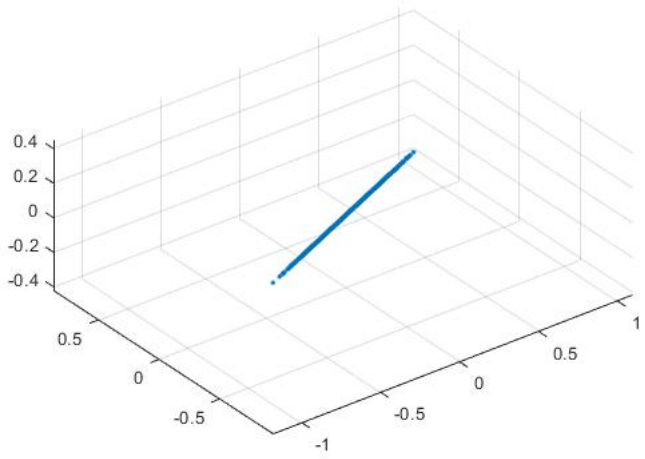

Figure 4: Scatter plot for $r = 1$

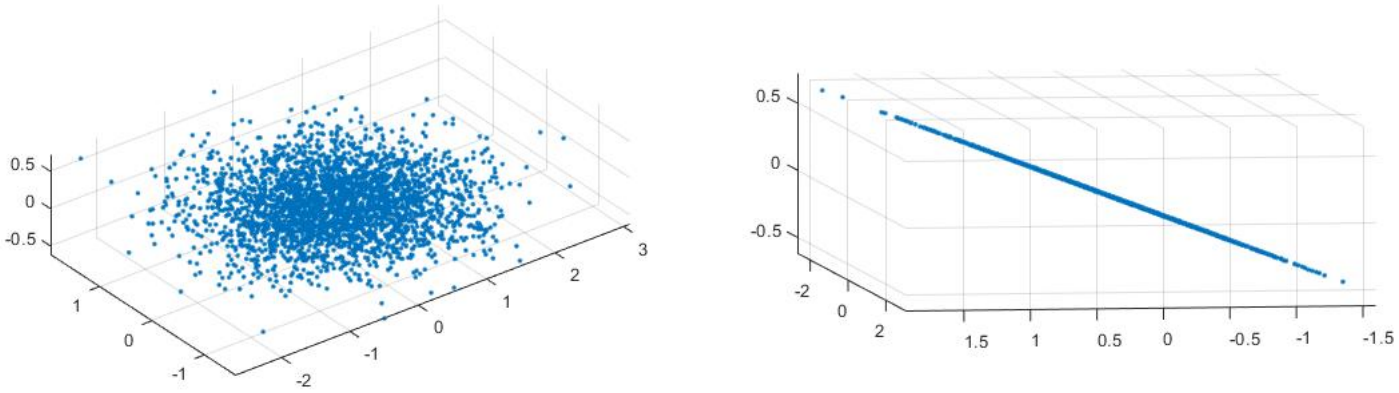

Figure 5: Scatter plot for $r = 2$ from two different views

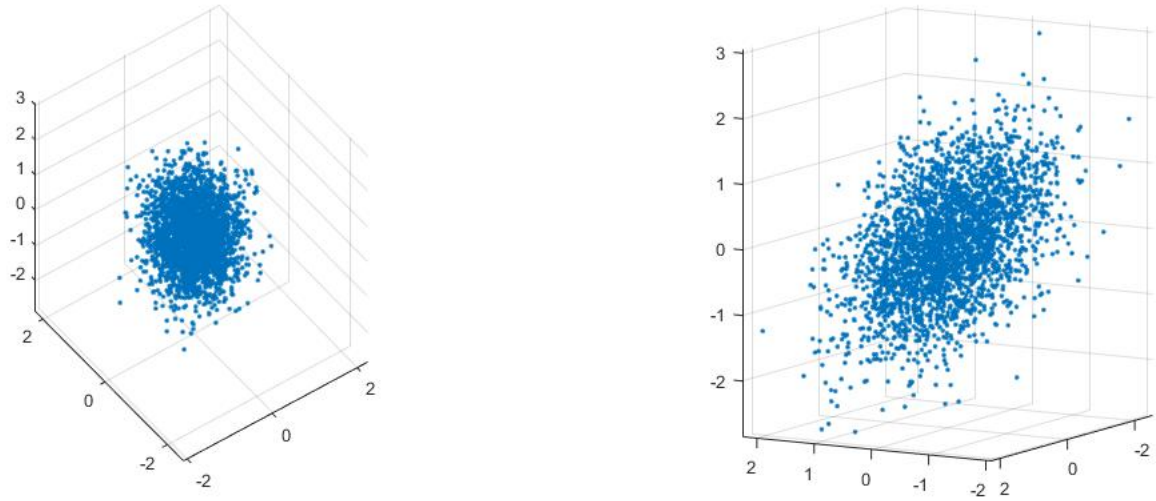

Figure 6: Scatter plot for $r = 3$ from two different views

4. Finally, modify the MATLAB code to simulate a case where the covariance matrix has rank 2, where the largest nonzero eigenvalue is equal to 1, and the other nonzero eigenvalue is equal to 0.01. Even though the covariance matrix has rank 2, what can you say about the approximate or effective dimension of the data in this case? Submit a representative plot of the data in this case, along with a written explanation of your observations.

Solution: As shown in Figure 7, though the structure is still a 2D plane as we expected from part 1 and 3, the shape is similar to a bold line. Setting the second nonzero value 0.01, which is close to zero, makes the probability that the points locate along that axis very small compare the axis with eigenvalue=1. Thus the shape is shrunk from a plane to a "bold" line.

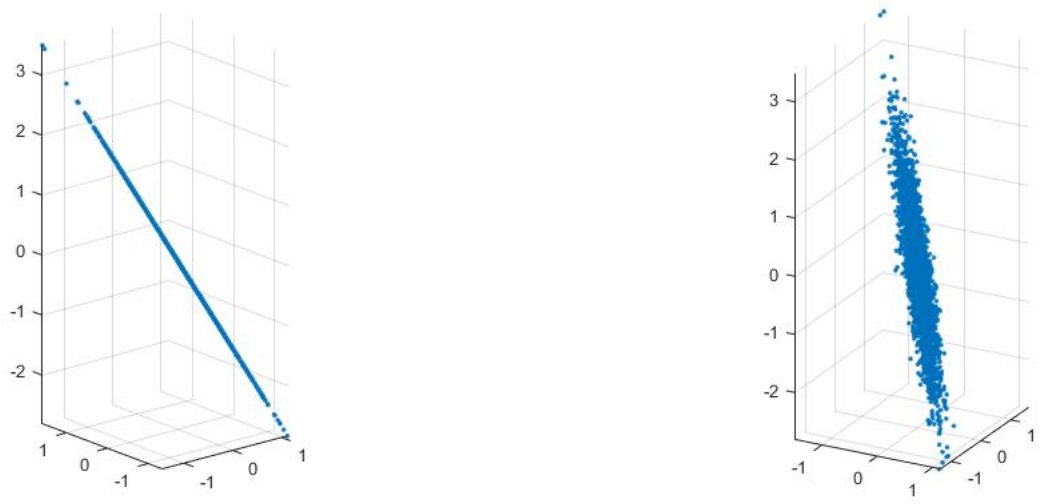

Figure 7: Scatter plot for $r = 2$ from two different views

2 Principal Component Analysis (PCA)

3. Submit figures showing the original images along with their 5 approximations, as well as a written discussion of your observations on how the quality of the approximation varies as r increases.

Solution: As the r increases, as shown in the figures below, the quality of the approximation improves. The idea of PCA idea is to represent high-dimensional data in a succinct way by projecting it into a lower dimensional subspace of estimators, with r ($r < p$) necessary degrees of freedom. The blur of the figures shows that the chosen degree of freedom is not sufficient to get a clear estimation.

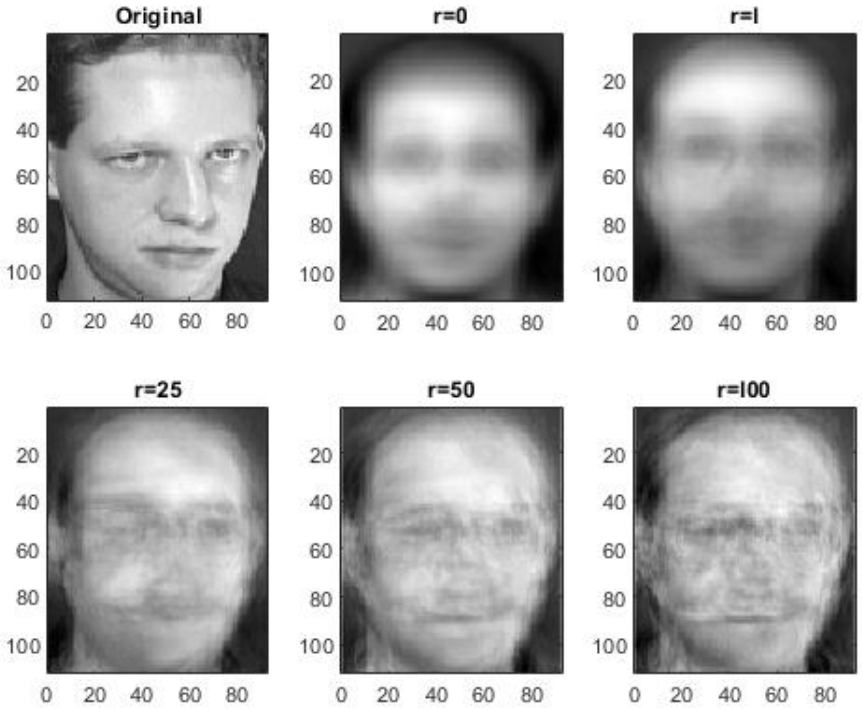

Figure 8: The first original images along with their 5 approximations

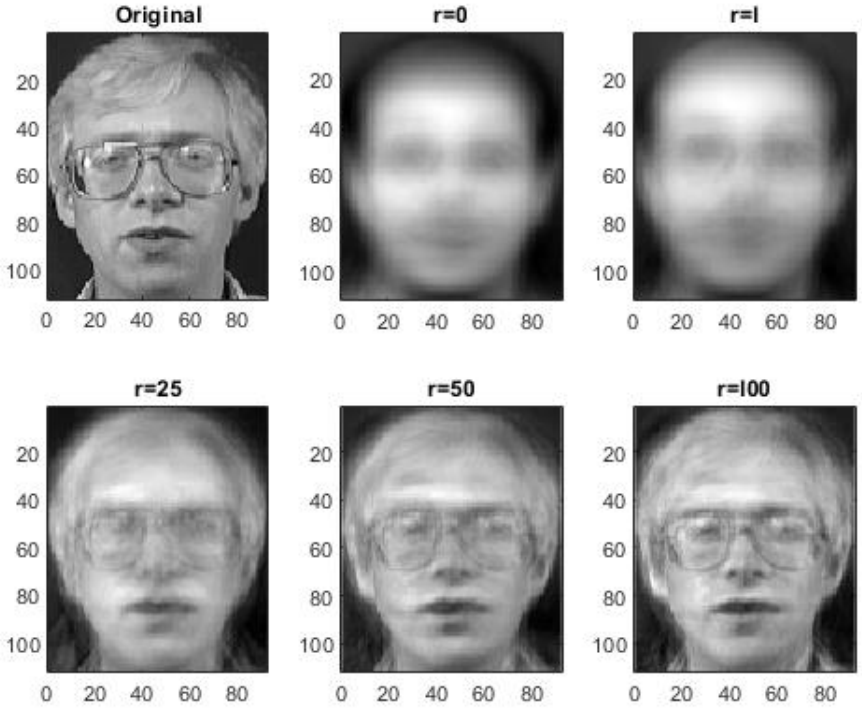

Figure 9: The second original images along with their 5 approximations

4. Submit this plot, along with a written discussion describing what you believe is the approximate effective dimension of the data (use what you learned in Part 1 of the project!)

Solution: From the conclusion from Q4 in Part 1, we know that for r = 1, the first principal component direction is the direction of maximum variation of the mean-centered data (or, the direction of maximum variance), the second principal component direction is orthogonal to the first, and captures the direction of the next most variation, and so on. The eigenvalues near zero indicate small relevance and thus not effective. Hence for this case, $r = 350$ is the approximate effective dimension as shown in Figure 10.

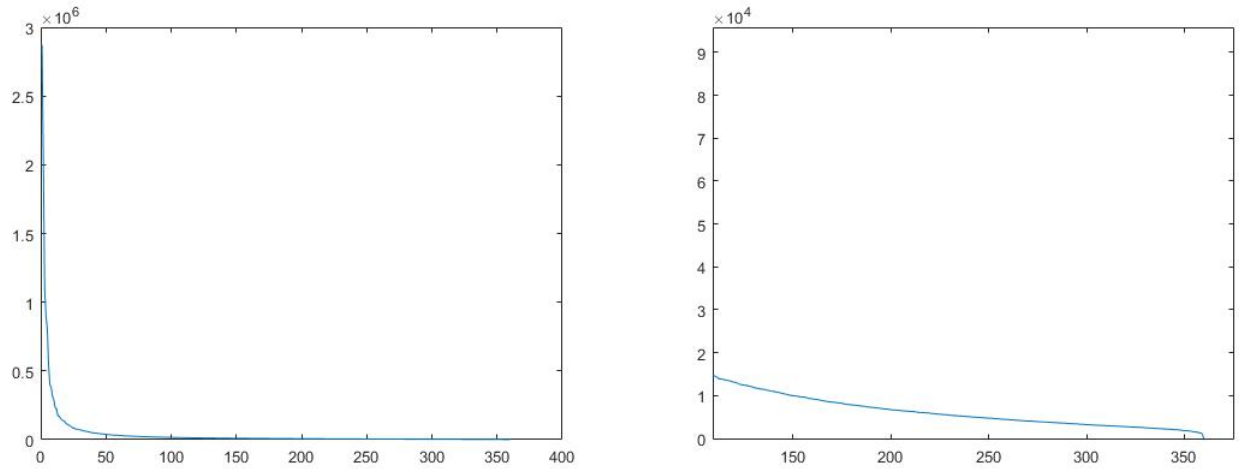

Figure 10: The plot of eigenvalues

3 Experimentation!

For this part, you will apply the PCA approach to a different data set of your choosing. Submit the plot of eigenvalues of the training data for your data set, and explain what you believe is an appropriate effective dimension (analogously to task 4 in Part 2 above). Also, submit a few plots showing approximations of some test data (points that must be omitted from the training set!) using the principal components you learned from your training data (analogously to task 3 in Part 2 above).

Solution: I chose the IMM Face Database[1], an annotated dataset of 240 face images, as my data set. Using the similar way that we extract data from images before ('readerCopy.m') and the

same PCA method('PCAtest.m'), the simulation results are shown below. The plot of eigenvalues shown as Figure 10 indicates the effective r value would be 200. As shown in the result in Figure 12 and Figure 13, as the r increases to 200, as shown in the figures below, the quality of the approximation improves.

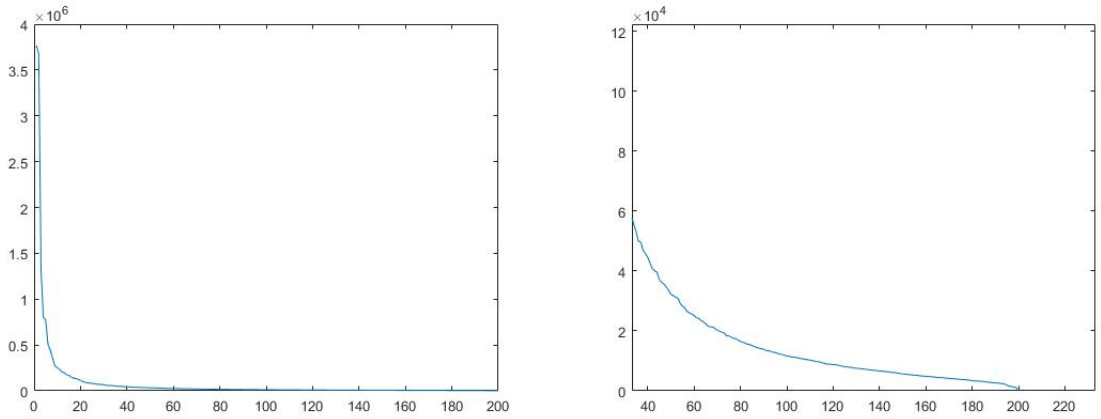

Figure 11: The plot of eigenvalues (the right view is zoomed-up)

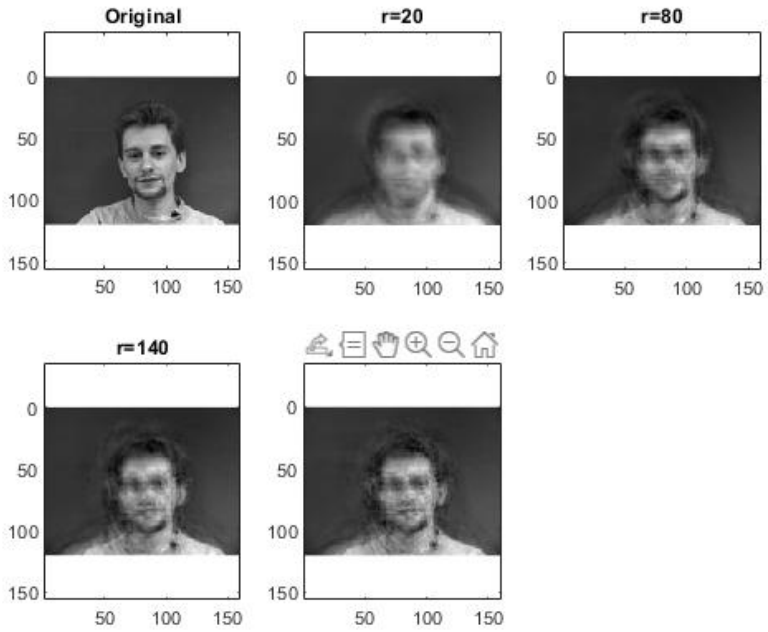

Figure 12: The first original images along with their 4 approximations

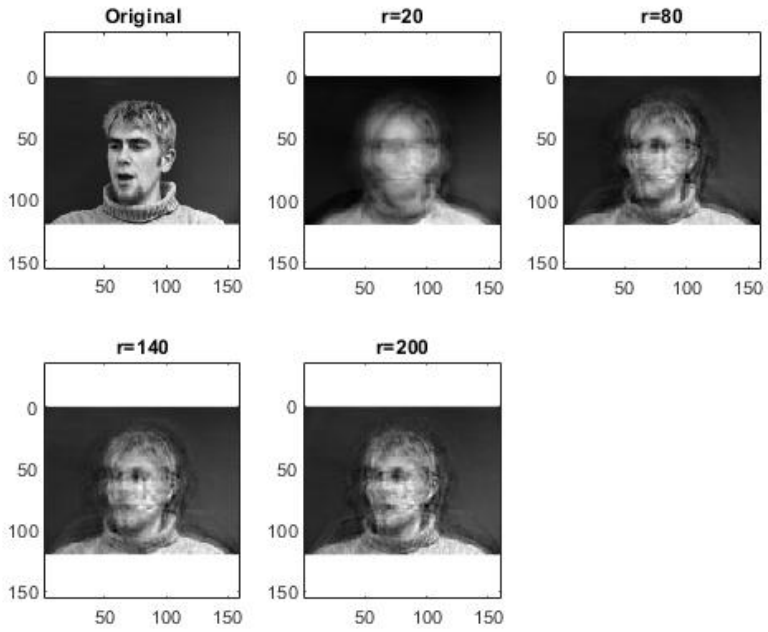

Figure 13: The second original images along with their 4 approximations

References

[1] M. M. Nordstrøm, M. Larsen, J. Sierakowski, and M. B. Stegmann. The IMM face database - an annotated dataset of 240 face images, may 2004.